DATA MINING FOR CLASSIFICATION OF HIGH VOLUME DENSE LiDAR DATA IN AN URBAN AREA

I. Chauhan [1]*, C. Brenner [2]

[1] Civil Engineering Department, G.B. Pant Institute of Engineering and Technology, Pauri Garhwal, Uttarakhand, India- inshu0302@gmail.com
[2] Institute of Cartography and GeoInformatics, Leibniz Universität Hannover, Germany - Claus.Brenner@ikg.uni- hannover.de

Commission V, SS: Emerging Trends in Geoinformatics

KEY WORDS: LiDAR, Urban Area, Classification, Data Mining, Random Forest, Machine Learning, Point Cloud

ABSTRACT:

3D LiDAR point cloud obtained from the laser scanner is too dense and contains millions of points with information. For such huge volume of data to be sorted, identified, validated and be used for prediction, data mining provides immense scope and has been used to achieve the same. Certain unique attributes were selected as an input for creating models through machine learning. Supervised models were thus built for prediction of classes through the available LiDAR data using random forest algorithm. The algorithm was chosen owing to its efficiency and accuracy over other data mining algorithms. The models created using random forest were then tested on an unclassified point cloud data of an urban area. The method shows promising results in terms of classification accuracy as overall accuracy of 91.71 % was achieved for pixel-based classification. The method also displays enhanced efficiency over common classification algorithms as the time taken to make predictions about the data is reduced considerably for a set of dense LiDAR data. This shows positive foresight of making use of data mining and machine learning to handle large volume of LiDAR data and can go a long way in augmenting efficient processing of LiDAR data.

1. INTRODUCTION

1.1 Classification of an urban scene

For urban development planning, forecast and simulation and for automatic positioning of vehicle, it has been desirable to extract terrain and building information in such an urban area (Shan and Sampath, 2007). An urban area consists of mixed features mainly like roads, trees, buildings, traffic, poles. The extraction of these features has been a difficult task due to the proximity of the various features that exist in an urban area.

Remote sensing has been the basic source of collecting the information on urban areas till date. At high spatial resolutions, the extraction of urban information is a challenging task in remote sensing. This is because the per-pixel methods that are commonly used are likely to fail due to their inability to capture the increased natural inconsistencies in the reflectance, and due to the reality, that each class category may contain several spatially adjacent pixels (Y.O. Ouma1, R. Tateishi1, J.T. and Sri-Sumanty, 2010).

However, deriving accurate, quantitative measures from remote sensing imagery over urban areas is still a research challenge due to the great spectral and spatial variability of the materials that compose urban land cover (Xian & Crane, 2005). When the sensor's instantaneous field-of-view (IFOV) has more than one land cover on the ground, there are mixed pixels in the remotely sensed imagery. The highly heterogeneous nature of urban surface materials creates problems at different spatial levels to classify urban areas easily and accurately.

1.2 LiDAR data

Laser scanning provides an efficient way to actively acquire accurate and dense 3D point clouds of object surfaces or environments which can overcome the problems faced in classification of urban areas. These 3D point clouds provide a good basis for rapid modelling in industrial automation, architecture, agriculture, construction or maintenance of tunnels and mines, facility management, and urban and regional planning (Elseberg et. al, 2011).

A LiDAR works basically through laser source which emits light. Then the light reflected by the objects is collected by the sensor which is then used to calculate ranges or distances to objects (NOAA, 2012). From this LiDAR survey, the datasets generated are of two types. One, scanned image of the area and second, information about the x, y and z coordinates of the scanned points. Eventually, the data used from LiDAR for classification is commonly in form of points or point cloud. The information about x, y and z of the points can be stored as text files; however, the size of these files may be huge.

The complexity and sheer volume of points in typical LiDAR dataset makes it difficult to work with (NOAA, 2012). Most scanners in use these days acquire data at astonishing rate of about 500,000 measurements/sec. Handling such dense and large amount of 3D data is difficult and challenging in real time situations. Classification of such an enormous amount of points remains a challenge throughout the scientific community. The research here aims at developing an efficient method to process and classify this high volume of LiDAR data points.

* Corresponding author

1.3 Data Mining

Classification is basically making predictions about unseen examples based on observations. An algorithm is such a set of course of action which helps in making predictions about the data. So, a classification algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations. One associated term with attaining information from a substantial quantity of data is data mining. Data mining is defined as a technique used in various domains to give meaning to the available data or specifically process of discovering patters in the data (Witten and Frank, 2005).

3D LiDAR data obtained from the laser scanner is too dense and contains millions of points with information. For such huge volume of data to be sorted, identified, validated and be used for prediction requires a technique like data mining. In data mining one of the most common tasks is to build models for the prediction of the class of an object based on its attributes. The above approach was tactfully adopted in this research work. Certain attributes in form of feature vector were delineated based on geometric neighbourhood in the LiDAR data after applying PCA. The feature vectors were taken as an input while applying random forest data mining algorithm to the acquired data set.

2. STUDY AREA AND DATA

The Mobile Mapping System was driven through the Oster Strasse which is a street of a stretch of 1.5 km in Hannover city, Germany (Figure 1). The Mobile Mapping System consists of the laser scanner, RIEGL VMX-250 (Figure 2) which is mounted on a Volkswagen Van. The scanner scans the area with low-noise, gapless 360° profiles at a measurement rate of 300,000 measurements/sec. and a scan rate up to 100 profiles/sec, for each of two scan heads. The raw LiDAR data is pre-processed into the two main data, one is yml file containing x,y,z co-ordinates and second is the intensity image (Figure 3) created using laser returns.

The yml file is essentially the raw scan data from one of the Riegl Scanners, processed in a way that for each turn of the scan head, a new column is started. The range and distance measurements of the scan head are transformed to XYZ coordinates using the position and orientation of the van (which in turn is taken from the built-in GPS/IMU) and all required transformations between scan head coordinate system and van (i.e., IMU) coordinate system. This transformation is done for every single scan point. The intensity image is taken from the "intensity" channel of the laser scanner, which is related to the amplitude of the backscattered laser pulse.

Figure 1: A google image of the urban scene taken as study area

Figure 2: RIEGL VMX-250, mounted on a van

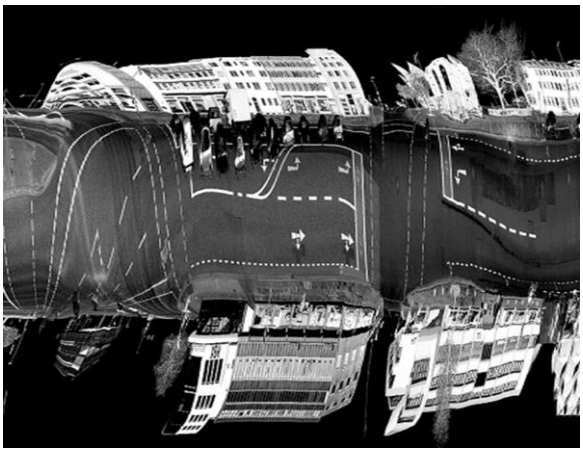

Figure 3: Intensity Image obtained from laser scanner

3. APPROACH & METHODOLOGY

The principal component analysis is performed to get a basic idea of planar, linear and volumetric surfaces in the 3D data. Based on PCA results and geometric properties, feature vector is developed for classification consisting of 7 attributes. The intensity image of the study area was used to classify the objects, colouring each class manually with a colour defined before. This is termed as hand classified image in the following sections. Data mining algorithms (using Weka user interface) were used for training data and models were saved. Lastly the models developed based on the feature vector are applied on the test data to get the final classification.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) has been used to obtain the geometric orientation of objects in a 3D point cloud in space. The local spatial point distribution over a neighbouring area is captured by the decomposition into principal components of the covariance matrix of 3D point position (Lalonde et. al, 2006). PCA approximates the spatial distribution of points in the neighbourhood by an ellipsoid with axis Vi and axis length σi = √λi (Monnier et al., 2012).

Principal component analysis is a transform of a given set of n input vectors (variables) with same length K formed in the n-dimensional $[x_1, x_2, \ldots\ldots x_n]^T$ vector into a vector y according to

$$y = A(x - m_x) \quad (1)$$

The symmetric positive definite covariance matrix for a set of N 3-D points (Lalonde, 2011) $\{X_i\} = \{(x_i, y_i, z_i)^T\}$ with $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is defined as

$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})\,(X_i - \bar{X})^T \quad (2)$$

The matrix is decomposed into principal components ordered by increasing eigenvalues. $\overrightarrow{e_0}, \overrightarrow{e_1}, \overrightarrow{e_2}$ are the eigenvectors corresponding to the eigenvalues $\lambda_0, \lambda_1, \lambda_2$ respectively, where $\lambda_0 \geq \lambda_1 \geq \lambda_2$. In case of scattered points, the eigenvalues are related as $\lambda_0 \approx \lambda_1 \approx \lambda_2$ and no dominant direction can be established. In case of a linear points, the principal direction will be tangent at the curve, so the relation of eigenvalues changes to $\lambda_0 \gg \lambda_1 \approx \lambda_2$. Finally, in the case of planar points, the principal direction is aligned with the surface normal with $\lambda_0 \approx \lambda_1 \geq \lambda_2$ and $\overrightarrow{e_0}, \overrightarrow{e_1}$ span the local plane of observations (Lam et al., 2011).

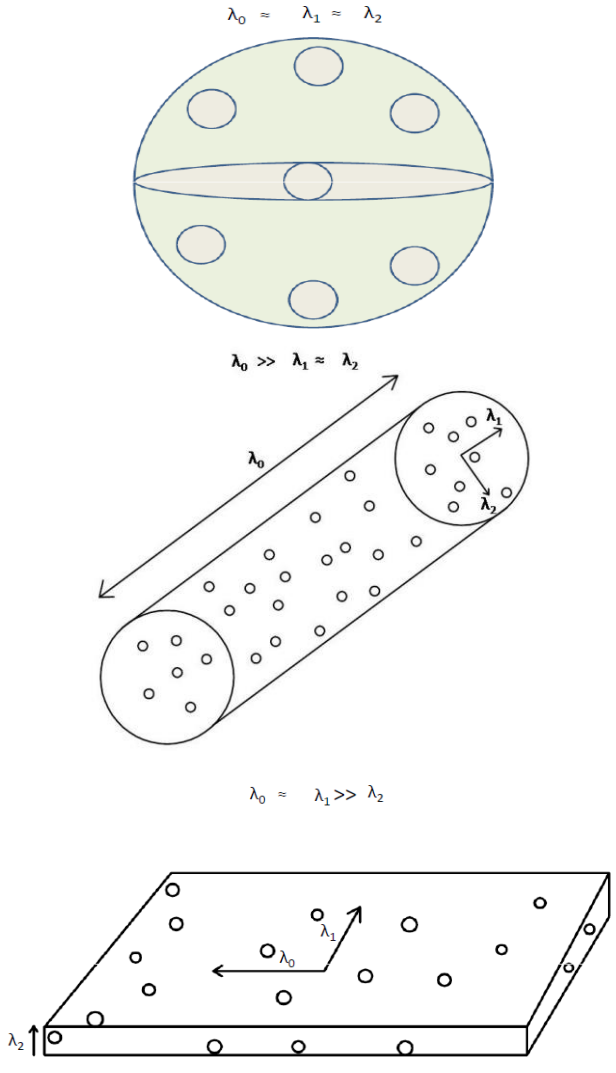

Figure 4: Representation of volumetric, linear, planar areas in a LiDAR data (Lalonde et. al, 2006)

In the given study area, trees represent scattered objects, poles and trunk of trees represent linear objects and building facades and pavement surfaces as planar objects. The PCA based classification just predicts the geometric condition of the objects in space but cannot identify which object is it.

3.2 Feature vector determination

Feature extraction can be viewed as finding a set of vectors that represent an observation while reducing the dimensionality. In pattern recognition, it is desirable to extract features that are focused on discriminating between classes. Although a reduction in dimensionality is desirable, the error increment due to the reduction in dimension must be without sacrificing the discriminative power of the classifiers (Benediktsson et al., 2003).

Geometry is a governing factor in processing of LiDAR data, so attributes were developed based on, in space geometric properties of various objects obtained from the LiDAR point cloud. For this a dynamic neighbourhood region of .5 m was defined. Based on the pixels contained by the region within the defined neighbourhood, the seven attributes were taken for classification. Together taken as feature vector these are summarized in Table 1.

S.No.	Attribute name	Attribute Interpretation
1.	Elevation of the points (z)	Different objects in a 3D environment have variable elevation but in general groups can be formed based on elevation
2.	Minimum Eigenvalue (λ_0)	Geometrically the areas with lowest eigenvalue will mean that this neighbourhood has minimum variation in any direction can be mainly planar surface
3.	Maximum Eigenvalue (λ_2)	Represents the most variation in data e.g. scattered points (trees or corners of buildings)
4.	Scalar product of up vector with normal vector	If the scalar product made by the normal vector with the up vector is small, it can be concluded that it's a road or ground surface otherwise if it makes a larger angle with the up vector then it can be a building facade
5.	Scalar product of vector perpendicular to up vector with normal vector	Represents the vector of the direction corresponding to maximum eigenvalue, helpful in distinguishing linear objects
6.	Ratio of maximum eigenvalue to the central eigenvalue	Represents the planar condition of points in space. Planar surfaces like building facades and roads can be differentiated from vegetation points and linear surfaces using this condition
7.	Ratio of maximum eigenvalue to minimum eigenvalue	Represents the volumetric condition of points in space, helps to strengthen the difference between trees and other objects in the urban scene

Table 1: Summary of feature vector to be used in data mining

Feature vector for each point are developed using python script as it can handle enormous data well.

3.3 Classification: Data Mining

There were seven classes identified from given urban scene which were to be extracted from the LiDAR data set. These were namely “tress”, “buildings”, “cars”, “road”, “trees’ branches and trucks”, “poles”, and “traffic lights”. The data contained in the yml file, consists of 6,204,909 points. Reading and classifying such enormous quantity of points at the same time is an exigent task. The yml file contains x, y, z co-ordinates of every point, which makes the data more extensive. This issue was overcome by using data mining, a number of algorithms like decisionstump, j48, random forest (Breiman 2001) and desicionrule were tried on a part of the data set. It was observed that random forest provided the best accuracy among all algorithms. Thus, random forest algorithm was used to create supervised models from the hand classified image (Figure 5) using $1/10^{th}$ part of the full data including the feature vector. This data for training was elected such that it covers some good quantity of pixels of all the classes.

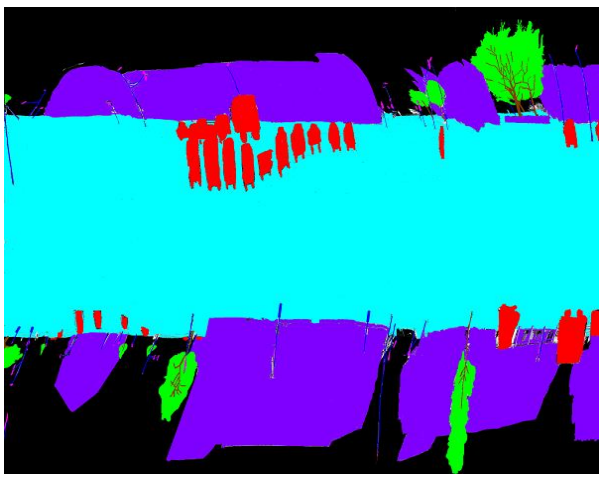

Figure 5: Hand classified image

The process involved in classifying the rest of the data is that with the help of statistical model developed using random forests, the feature vector of this data is coordinated according to it and the decision to generate new class is made.

4. RESULTS AND DISCUSSION

4.1 Results

Application of only PCA showed that roads, parts of buildings as well as some parts of thick tree trunks are classified as planes. Tree leaves and corners of windows are classified as scattered, volumetric points. Rest of the points are classified as linear representation. Some parts of facades (corners of windows and buildings) are classified as trees, also some parts of tree trunks classified as facades, and many such inaccuracies (Figure 6). Also, the PCA classifies some surfaces as planar which in fact may be a road or a building facade or a car surface. To further improve classification, feature vectors for a dynamic neighbourhood of .5 m are developed and are assigned into the training model. Using python script, running the program and classification took more than 12 hours and sometimes the well configured computer would hang owing to large amount of data. To reduce the time as well as to increase the accuracy, data mining proved to be efficient. After applying random forest algorithm, the time taken to classify the urban scene reduced to almost half. Random forest ran efficiently on the large database and handled thousand of input variables without variable deletion which resulted in excellent accuracy of 91.71%. Most classes were correctly extracted (Figure 7).

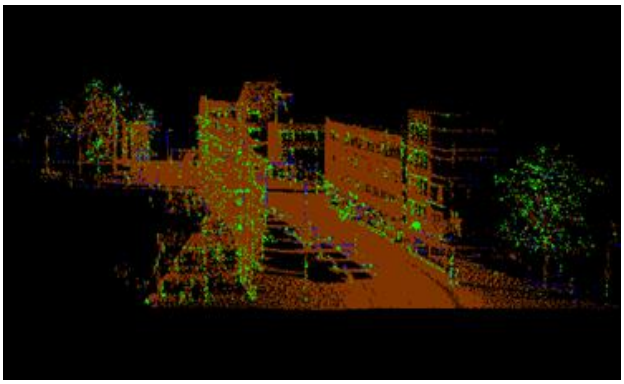

Figure 6: Cross section of the urban area classified after PCA

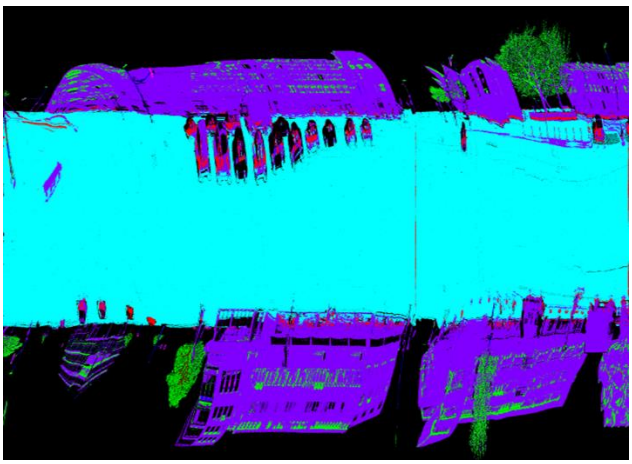

Figure 7: Accurately classified image with aid from data mining

It can be observed from above figure that trees and roads are classified very accurately through pixel-wise supervised classification while there are some mixed pixels in buildings, poles, cars, trees trunks and branches and traffic lights.

4.2 Discussion

The complexities of an urban environment make it difficult to apply simple, conventional methods of surveying to make 3D models of a city easily. LiDAR point cloud is proving to be of great importance in this scenario, if its limitation of being colossal can be reduced through use of modern tools like data mining. The integration of geometric features through PCA and machine learning through random forest algorithm for pixel-based classification was achieved successfully through the proposed methodology. The objective of classifying a complex urban environment with different classes has been attained satisfactorily. The LiDAR data in-spite of being in such enormous quantity has been handled well by python as well as Weka (data mining tool) while performing analysis for classification.

This shows positive foresight of making use of data mining and machine learning to handle large volume of LiDAR data and can go a long way in augmenting efficient processing of LiDAR point cloud. The classification results obtained from this study can be applied to variety of field like urban planning, infrastructure planning, creating 3D models of the city, in

automatic navigation of cars for positioning of vehicles. This study is not without its limitations. Though the time of classification has been reduced considerably, real time automated classification will take a long way to go. Also, the work here tried limited algorithms (as available in Weka) with respect to data mining, more of them can be explored and greater efficiency can be achieved.

REFERENCES

Benediktsson, J.A., and Syeinsson, J.R., 1997. Feature extraction for multisource data classification with artificial neural networks. In: International Journal of Remote Sensing, Vol. 18(4), pp. 727-740, https://doi.org/10.1080/014311697218728.

Breiman, L., 2001. Random Forests. In: Machine Learning, Vol. 45, pp. 5-32.

Elseberg, J., Borrmann, D., and Nuchter, A., 2011. Efficient processing of large 3d point clouds. In: Information, Communication and Automation Technologies (ICAT), 2011 XXIII International Symposium, IEEE 2011, pp. 1-7.

Lalonde, J.F., Vandapel, N., Huber, D.F., and Hebert, M., 2006. Natural terrain classification using three-dimensional ladar data for ground robot mobility. In: Journal of field robotics, Vol. 23(10), pp. 839-861.

Lam, J., Kusevic, K., Mrstik, R., Harrap, P., and Greenspan, M., 2011. Urban scene extraction from mobile ground based lidar data. In: Proc. 3DPVT, Vol. 2010, pp. 478-486.

Monnier, F., Vallet, B., and Soheilian, B., 2012. Trees detection from laser point clouds acquired in dense urban areas by a mobile mapping system, https://doi.org/10.5194/isprsannals-I-3-245-2012.

National Oeanic and Atmospheric Administration (NOAA) Coastal Services Center, 2012. Lidar 101: An Introduction to Lidar Technology, Data and Application (Revised).

Ouma, Y.O., Tateishi, R., and Sri-Sumantyo, J.T., 2010. Urban features recognition multi-spectral satellite imagery: a micro-macro texture determination and integration framework. In: Image Processing, IET, Vol. 4(4), pp. 235-254.

Shan, J., and Sampath, A., 2007. Urban terrain and building extraction from airborne LIDAR data. CRC Press, Boca Raton (USA), pp. 21-42.

Witten, I.H., & Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Second Edition.

Xian, G., and Crane, M., 2005. Assessments of urban growth in the Tampa Bay watershed using remote sensing data. In: Remote Sensing of Environment, Vol. 97(2), pp. 203-215.