Structured and Unstructured Machine Learning for Crowdsourced Spatial Data

Musfira Jilani MUSFIRA.JILANI@UCDCONNECT.IE
University College Dublin, Dublin, Ireland

Padraig Corcoran CorcoranP@cardiff.ac.uk
Cardiff University, Wales, UK

Michela Bertolotto MICHELA.BERTOLOTTO@UCD.IE
University College Dublin, Dublin, Ireland

Abstract

Recent years have seen a significant increase in the number of applications requiring accurate and up-to-date spatial data. In this context crowdsourced maps such as OpenStreetMap (OSM) have the potential to provide a free and timely representation of our world. However, one factor that negatively influences the proliferation of these maps is the uncertainty about their data quality. This paper presents structured and unstructured machine learning methods to automatically assess and improve the semantic quality of streets in the OSM database.

1. Introduction

The need and importance of accurate spatial data has never been greater. However, given the vastness and dynamics of our world, mapping is an expensive process and obtaining timely and accurate spatial information is a challenging process. In this context, crowdsourced maps such as OpenStreetMap (OSM), have the potential to provide a free and up-to-date representation of our world. However, similar to most other crowdsourced platforms, concerns exist about the quality of such maps. In this paper, we present a novel approach for automatically assessing and improving the quality of crowdsourced spatial data. Specifically, we focus on the semantic type quality of streets in OSM where by semantic type we mean the class of a street such as motorway, residential, etc. Toward the goal of assessing and possibly improving the semantic quality of streets in the OSM database, we develop structured and unstructured (classical) machine learning models that can learn from the geometrical and topological features of a street network the semantics of its constituent streets. In addition, the structured model can also exploit the inherent spatial relationships between various streets in a given network.

2. Related Work

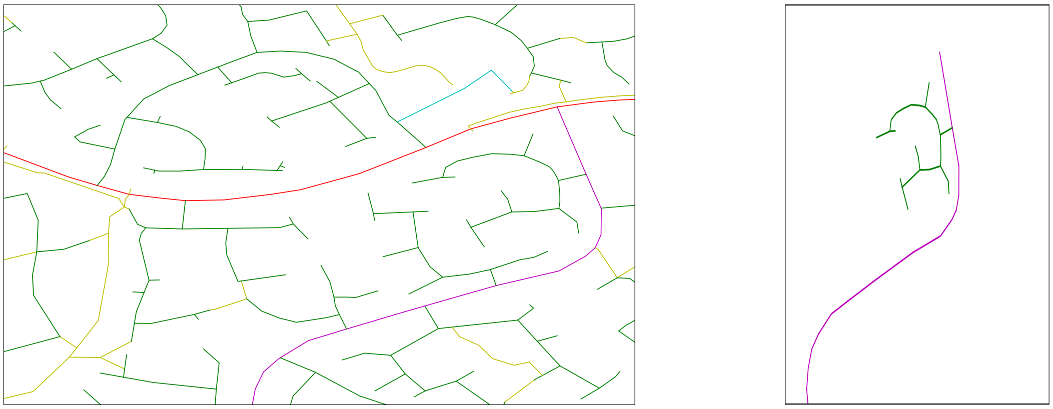
Most of the current data quality assessment techniques for OSM require referencing to commercial or official maps (Girres & Touya, 2010),(Haklay, 2010). However, we argue that this process of comparing a crowdsourced (heterogenous) database with authoritative maps is ineffective. Instead we focus on the use of machine learning techniques that can enable us to assess and possibly improve the data quality of crowdsourced maps without referencing to external databases.

3. Methodology

Toward the goal of quality analysis and quality assessment of streets in OSM data, we propose a machine learning based methodology. The key idea is that if we can learn the semantic type of streets, we can predict them and make use of these predictions to assess the accuracy of existing semantics and also possibly fill the missing semantics.

A street network consists of several street segments where a street segment can be defined as the section of a street between two intersections, or between a dead-end and an intersection. A close observation of a street network suggests that the semantic type information of

Preliminary work. Under review by the Indian Workshop on Machine Learning (*i*WML). Do not distribute.



(a)

(b)

Figure 1. A section of street network from OSM London. (a) Five different street types can be seen: secondary(red), tertiary(magenta), residential(green), footway(yellow), service(cyan).(b) Two connected sets of street segments, one each from tertiary and residential types respectively have been redrawn for comparison.

streets is implicitly contained within connected sets of street segments having uniform semantic type. Figure 1(a) shows a small part of London OSM street network consisting of five semantic types of streets each represented using different colours. Various simple distinguishing features of these semantic types are evident. For example, in Figure 1(b), a comparison of two connected sets of street segments belonging to two different street types: tertiary (magenta) and residential (green) is shown. The connected set of segments corresponding to the residential type has more dead-ends as compared to the set corresponding to the tertiary type. Similarly, the overall linearity for the tertiary set of street segments is much higher than the residential set where we define linearity as the degree to which the street segments in question have a shape similar to that of a straight line. Furthermore, the residential set is only connected to the tertiary set whereas the tertiary set is connected to both the residential and secondary sets of street segments. This example suggests that the features of connected sets of segments of uniform type vary as a function of type.

It follows from the above discussion that if we can measure the features enabling distinction between various semantic types of streets, we can subsequently use them for learning the semantic types of streets. However, these features are a function of a connected set of street segments where all elements of a particular set have the same semantic type. In the following sections we describe the methods used for identifying such sets of connected street segments, measuring their features, and subsequently using these features to learn the semantic types of streets in a crowdsourced spatial database such as OSM.

3.1. Data Representation

Appropriate data representation is a fundamental step toward useful knowledge discovery. Therefore, as a first step a novel multi-granular graph-based street network representation system is developed. All streets having same name and same semantic type correspond to a single node in a multi-granular graph. Such a representation makes the various features of a street explicit as opposed to implicit. More details of the multi-granular representation system can be found in (Jilani et al., 2013).

3.2. Feature Extraction

Using the multi-granular representation obtained above we extract several topological and geometrical features of streets such as length, linearity, number of dead-ends, number of intersections, semantic types of adjacent streets (using a BoW model), node degree, and betweenness centrality. These features are mainly inspired from the domains of computer vision, graph theory, and street network analysis.

3.3. Unstructured Learning

Next, we develop an unstructured (or classical) supervised machine learning model to learn the various

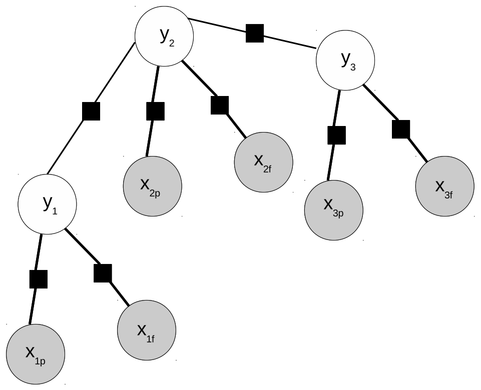


Figure 2. Graphical modelling of the street network. Label y_1 for the street x_1 depends on its observed features x_{1p} and x_{1f} by two unary potentials and on its neighbouring street's label y_2 by a pairwise potential.

semantics types of streets in the OSM database. The development of this model involves assessing the performance of the commonly used machine learning classifiers such as naive bayes, SVM, neural networks, and random forests in terms of their generalization performance on test data. More details on the implementation of the unstructured learning of the problem can be found in (Jilani et al., 2014).

3.4. Structured Learning

A street network is a structured input as it consists of several streets, where not only the streets themselves contain information such as geometry, but also the way in which the streets are connected to each other is important. For such a structured input, we obtain a structured output of semantic types of streets over all the streets in the network. We exploit the Conditional Random Field (CRF) framework for performing structured prediction. The CRF framework allows us to leverage prior knowledge available to us in the form of crowdsourced semantics, the geometrical and topological features of individual streets, and the contextual (structural) relationships between various streets into a single unified model.

Suppose we have a street network consisting of N streets $\mathbf{x} = \{x_1, x_2, \ldots, x_N\} \in \mathcal{X}$. Our goal is to predict a street type labelling $\mathbf{y} = \{y_1, y_2, \ldots, y_N\} \in \mathcal{Y}$ for these streets. Figure 2 shows our graphical model. x_{ip}, x_{if} are the observed features of a street x. More specifically, x_{ip} represents the initial crowd sourced labels or priors and x_{if} represents the geometric features. Toward the goal of jointly learning the labels $\mathbf{y}$, our model maximizes the conditional probability of $\mathbf{y}$ given $\mathbf{x}$:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}|\mathbf{x};\mathbf{w}) \tag{1}$$

where, using the Hammersley Clifford Theorem (Koller & Friedman, 2009), $P(\mathbf{y}|\mathbf{x};\mathbf{w})$ may be expressed as:

$$P(\mathbf{y}|\mathbf{x};\mathbf{w}) = \frac{exp(\psi(\mathbf{y},\mathbf{x};\mathbf{w}))}{\sum_{\mathbf{y'}\in\mathcal{Y}} exp(\psi(\mathbf{y'},\mathbf{x};\mathbf{w}))} \tag{2}$$

where $\psi(\mathbf{y},\mathbf{x};\mathbf{w})$ is a potential function measuring the compatibility between the output labels $\mathbf{y}$ given the input $\mathbf{x}$ and $\mathbf{w}$ are the model parameters. This potential function consists of three types of potentials corresponding to the three properties described previously: two unary potentials and one pairwise potential. More specifically, one unary potential measures the compatibility between the label y given the initial crowdsourced prior for the label x_p. Another unary potential measures the compatibility between the label y given the geometric features of the street x_f. Finally, the pairwise potential models the compatibility between neighbouring street labels. The potential $\psi(\mathbf{y},\mathbf{x};\mathbf{w})$ is linear in these basis potentials and parameters and can be compactly written as:

$$\psi(\mathbf{x},\mathbf{y};\mathbf{w}) = \mathbf{w}^T\Psi(\mathbf{x},\mathbf{y}) \tag{3}$$

We use a max-margin approach for determining the weight vector $\mathbf{w}$ from training examples $(x_1, y_1), ..., (x_n, y_n)$. This involves minimization of the constrained quadratic optimization problem of Equation 4 (Tsochantaridis et al., 2004), (Joachims et al., 2009).

$$\begin{aligned} &\min_{\mathbf{w},\xi\geq 0} \frac{1}{2}w^T w + C\xi \\ &s.t.\ \forall(\bar{y}_1,\ldots,\bar{y}_n) \in \mathcal{Y}^n : \\ &\frac{1}{n}w^T\sum_{i=1}^{n}[\Psi(x_i,y_i) - \Psi(x_i,\bar{y}_i)] \geq \Delta(y_i,\bar{y}_i) - \xi \end{aligned} \tag{4}$$

The term ξ is a slack variable while $w^T w$ represents the size of the margin dividing the training examples. The term C balances these two terms in the objective. The term $\Delta(y_i,\bar{y}_i)$ is a function which measures the loss associated with a labelling $\bar{y}_i$ is the true labelling is y_i. In this work the Hamming loss[1] is used. In-

[1] Hamming loss computes the hamming distance between the true and predicted labellings.

tuitively, the above optimization problem models the fact that we wish to minimize the number of incorrect classifications while maximizing the margin.

A fusion moves approach was used for inferring the street labellings. More details on the structured learning of the problem can be found in (Corcoran et al., 2015) and (Jilani et al., 2016).

4. Results and Discussion

We trained and tested our models on two non-overalapping regions from OSM London database. All 19 popular semantic types of streets used in OSM database for classifying a street were considered. An overall classification accuracy of 55.95% was obtained using the unstructured learning model (random forest). This accuracy increased to 84.75% when structured learning framework was used. Clearly, and naturally the structured learning framework outperforms the unstructured learning performance as it exploits the inherent structure in street networks. To the best of our knowledge, this is the first time that a structured learning framework has been used in the context of crowdsourced spatial data.

In this work, we considered all the 19 popular semantic types of streets used for classifying a street network. However, such a classification of street network is too fine-grained when compared with the commonly used and understood street network classifications where a street network is usually classified into 4-10 semantic types. While a fine-grained classification of street network may be ambiguous and less understood by common people, in our experiments also we observed that while the classification accuracy for certain types of streets such as residential and motorways is quite high, there are some overlapping street types such as footways and pedestrian which are difficult to learn.

In future we propose the development of a multi-layer conditional random field based model for simultaneously learning both the fine-grained (19) and coarse-grained (4-10) semantic types of streets. In addition, the models developed in this paper will also be extended to other map objects such as buildings, Points of Interests (PoIs), etc.

Dual Submissions

The work presented in this paper is a summary of the work already published at the following venues:

1. 23rd ACM SIGSPATIAL Conference, USA, 2015
2. 22nd ACM SIGSPATIAL Conference, USA, 2014
3. IWCTS, 22nd ACM SIGSPATIAL, USA, 2013
4. Intelligent Systems, Technologies, and Applications, Springer, 2016

In addition, a revised version of this paper will also be submitted toECML-PKDD, Nectar Track, Italy, 2016.

References

Corcoran, Padraig, Jilani, Musfira, Mooney, Peter, and Bertolotto, Michela. Inferring semantics from geometry: the case of street networks. In *Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in GIS*. ACM, 2015.

Girres, Jean Franois and Touya, Guillaume. Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14(4):435–459, 2010.

Haklay, Mordichai. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning*, 37(4):682–703, 2010.

Jilani, Musfira, Corcoran, Padraig, and Bertolotto, Michela. Multi-granular street network representation towards quality assessment of openstreetmap data. In *Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science*. ACM, 2013.

Jilani, Musfira, Corcoran, Padraig, and Bertolotto, Michela. Automated highway tag assessment of openstreetmap road networks. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in GIS*. ACM, 2014.

Jilani, Musfira, Corcoran, Padraig, and Bertolotto, Michela. Probabilistic graphical modelling for semantic labelling of crowdsourced map data. In *Intelligent Systems Technologies and Applications*. Springer, 2016.

Joachims, Thorsten, Finley, Thomas, and Yu, Chun-Nam John. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

Koller, Daphne and Friedman, Nir. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

Tsochantaridis, Ioannis, Hofmann, Thomas, Joachims, Thorsten, and Altun, Yasemin. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 104. ACM, 2004.