Machine Learning for Crowdsourced Spatial Data

Musfira Jilani[1(✉)], Padraig Corcoran[2], and Michela Bertolotto[1]

[1] School of Computer Science, University College Dublin, Dublin, Ireland
musfira.jilani@ucdconnect.ie, michela.bertolotto@ucd.ie
[2] School of Computer Science, Cardiff University, Cardiff, UK
CorcoranP@cardiff.ac.uk

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.

Keywords: Probabilistic graphical modelling · Crowdsourced spatial data · Street networks · Semantics

1 Introduction

We live in an age where the demand for accurate and up-to-date spatial data has never been greater. However, obtaining and maintaining such spatial databases is a challenging and expensive task. In this context, crowdsourced maps such as OpenStreetMap (OSM)[1] can be a viable solution for obtaining a free and up-to-date representation of our world. While an extensive number of applications have been developed around OSM, concerns exist regarding the quality of OSM data. The predominant method for assessing OSM data quality is based on comparing the OSM data with some form of authoritative maps such as the Ordnance Survey UK [2], Google Maps, etc. However, we argue that this process of comparing a crowdsourced (heterogenous) database with authoritative maps is ineffective. Instead we propose the use of machine learning techniques for assessing and possibly improving the data quality of crowdsourced maps without referencing to external repositories.

Specifically, in this paper we focus on the semantic type quality of streets in the OSM where semantic type refers to the class of a street such as motorway, pedestrian, etc. We hypothesize that the semantic types of streets are a

[1] The OSM project was started in 2004 with a goal of creating a free and editable map of the entire world. www.openstreetmap.org.

B. Berendt et al. (Eds.): ECML PKDD 2016, Part III, LNAI 9853, pp. 294–297, 2016.
DOI: 10.1007/978-3-319-46131-1_38

function of their geometrical and topological features and develop structured and unstructured machine learning models that can learn the semantic types of streets given such features. Interestingly, the structured learning models can also exploit the inherent spatial relationships within a street network.

2 Methodology

2.1 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 [3].

2.2 Feature Extraction

Several topological and geometrical features of streets were extracted using the multi-granular street network representation obtained above. These include length, linearity, number of dead-ends, number of intersections, semantic types of adjacent streets (using a BoW model), node degree, and betweenness centrality.

2.3 Unstructured Learning

Next, we develop an unstructured (or classical) supervised machine learning model to learn the various 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 [4].

2.4 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.

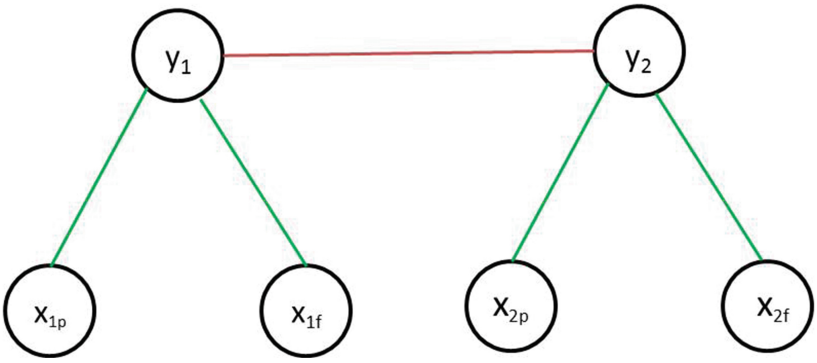

Fig. 1. Street network represented as a graphical model. x are the observed variables corresponding to the streets in the network and y are the labelling we want to infer. The green lines correspond to the unary potentials in the model and the red line to the pairwise potential.

Suppose we have a street network consisting of N streets $\mathbf{x} = \{x_1, x_2, \ldots, x_N\} \in \mathcal{X}$ and our goal is to predict the semantic type labellings $\mathbf{y} = \{y_1, y_2, \ldots, y_N\} \in \mathcal{Y}$ for these streets. Figure 1 shows our representation of such a street network as a graphical model where x_{ip} corresponds to the initial crowd sourced labels or priors and x_{if} corresponds to the geometric and topological features. Toward the goal of jointly learning the semantic type labelling $\mathbf{y}$, our model maximizes the conditional probability of $\mathbf{y}$ given $\mathbf{x}$ [6]:

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

We use a max-margin approach for determining the model parameters w and a fusion moves approach for inferring the street labellings. More details on the structured learning of the problem can be found in [1,5].

3 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. 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.

4 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. Intelligent Systems, Technologies, and Applications, Springer, 2016
4. UL-NUIG Research Day, 2016.
5. Related version submitted to the Indian Workshop on Machine Learning, 2016.

Acknowledgments. This work is supported by the Irish Research Council through the Embark Postgraduate Scholarship Scheme 2012.

References

1. Corcoran, P., Jilani, M., Mooney, P., Bertolotto, M.: Inferring semantics from geometry: the case of street networks. In: Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in GIS. ACM (2015)
2. Haklay, M.: How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. Environ. Plann. **37**(4), 682–703 (2010)
3. Jilani, M., Corcoran, P., Bertolotto, M.: 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)
4. Jilani, M., Corcoran, P., Bertolotto, M.: Automated highway tag assessment of openstreetmap road networks. In: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in GIS. ACM (2014)
5. Jilani, M., Corcoran, P., Bertolotto, M.: Probabilistic graphical modelling for semantic labelling of crowdsourced map data. In: Jilani, M., Corcoran, P., Bertolotto, M. (eds.) Intelligent Systems Technologies and Applications. AISC, vol. 385, pp. 213–224. Springer, Heidelberg (2016)
6. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT press, Cambridge (2009)