Semantic Reasoning for Smog Disaster Analysis

Jiaoyan Chen[1], Huajun Chen[1], and Jeff Z. Pan[2]

[1] College of Computer Science, Zhejiang University
{jiaoyanchen,huajunsir}@zju.edu.cn
[2] Department of Computer Science, The University of Aberdeen
{jeff.z.pan}@abdn.ac.uk

Abstract. Smog disaster is a severe global problem. Although it has been investigated for decades in environmental sciences, the analysis of smog data recently becomes an open problem in fields like big data and artificial intelligence. In this paper, we present our study of utilizing semantic reasoning techniques for accurate and explanatory smog disaster prediction. To this end, we enriched the smog data streams with background knowledge by ontology modeling, inferred underlying knowledge like semantic assertions and rules, built consistent prediction models by embedding the knowledge (i.e., assertions and rules) in machine learning algorithms, and finally provided explanations by rule-based reasoning.

Keywords: Smog Disaster, Semantic Reasoning, Ontology, OWL, Rule

1 Introduction

Smog disaster is a kind of severe air pollution event that negatively influences people's health and damages the environment[8]. In the past decades, it has attacked a large part of the population, especially in the fast developing economies like China and India[4]. To deal with smog disasters, they are widely studied in environment sciences with domain methods, e.g., chemical model and satellite remote sensing. In these studies, prediction of air pollutants is one of the most important problem because of its significance in real world applications.

With widely deployed physical sensors and big urban data, predictive analytics for smog disasters becomes an open reseach problem in the communities of data mining and machine learning[1, 2, 10, 11]. For example, the study U-Air[10] predicted the air pollution index for those urban areas where there are no air quality stations using correlation analysis, feature extraction and multi-view learning. Different from those in environment sciences, these studies model the prediction problem in perspectives of data science and artificial intelligence. However, they apply the background knowledge with manual exploratory analysis and feature engineering, ignoring knowledge representation and reasoning. This disables their capability of automatically incorporating the underlying knowledge with the prediction model, which limits their generalization to other contexts. Meanwhile, the pure machine learning based prediction models usually lack of explanation to the results.

On the other hand, semantic reasoning has recently been applied for the predictive analytics of spatio-temporal data[3, 5–7]. They implemented some widely used prediction techniques e.g., association rule mining and auto-correlation analysis on the semantic enriched data, thus utilizing background knowledge and reasoning for semantic enhanced prediction. For example, in the study of semantic traffic analytics[5], the researchers (1) interpreted the traffic related time-series into ontology stream, (2) inferred assertions and axioms for each portion, also known as ontology stream snapshot, (3) calculated the auto-correlation across snapshots, (4) mined semantic rules from snapshots that are semantically similar to the testing snapshot, and finally (5) predictively inferred the traffic congestion status with explanations.

In this study, we aim at bridging the gap between semantic reasoning and machine learning in the context of predictive smog disaster analysis. To this end, we first semantically enhanced the smog related time-series by modeling the domain knowledge with Web Ontology Language (OWL), and then inferred the underlying knowledge i.e., assertions and rules. We finally embedded these knowledge into basic machine learning algorithms to realize consistent sampling and automatic feature extraction. In brief, this study contributes to both application and methology: (i) it builds a more accurate and explanatory prediction model for smog disasters; (ii) it proposes a framework for incorporating semantic web techniques with machine learning algorithms, thus enhancing traditional prediction models with knowledge representation and reasoning.

2 Context

Multiple heterogenous time-series observed from both physical sensors and a Chinese microblogging website, also known as Sina Weibo are used for this study as shown in Table 1. All the records are tagged with geographical position i.e., latitude and longitude. The application aims at predicting a position's air pollution after a period of time e.g., 12 hours and 24 hours with all the current observations. We model the problem as a classification problem, where air pollution status is divided into 6 ranges (i.e., Good, Moderate, Unhealthy, Very Unhealthy, Hazardous and Emergent) according to a US standard based on AQI (Air Quality Index) metric and air pollution's health impact.

Datasets	Record #	Stream #	Coverage	Datasets	Record #	Coverage
air quality	∼78.53M	9	945 stations in 190 cities	POI	∼23K	Beijing & Shanghai
meteorology	∼150.1M	11		checkin	∼2M	
weather forecast	∼101.2M	7		tweet	∼23.36M	

Table 1. Details of the datasets with a time span from May 2013 to December 2014.

3 Framework

Fig.1 presents the high-level technical framework for predictive smog disaster analysis using both semantic web and machine learning (ML) techniques. To get accurate and explanatory prediction results, we need to (1) process the tweets (e.g., text sentiment analysis), checking and POI records to calculate index of those social factors that may influence air pollution[1], (2) model the background

knowledge with OWL2 ontology (i.e., TBox) using description logic fragment $\mathcal{ALC}$ and transform the time-series into streaming facts (i.e., ABox), (3) infer underlying assertions through entailments with streaming reasoner e.g., TrOWL streaming[9], (4) mine SWRL (Semantic Web Rule Language) rules across snapshots and calculate their confidence and support[5], (5)(6)(7) embed facts, assertions and rules as consistent vectors and feature vectors, with both of which ML models are trained by Stochastic Gradient Descent (SGD) algorithm, (8) apply the SWRL rules in reasoning with the matched rules being the explanations, and finally (9) ensemble the results.

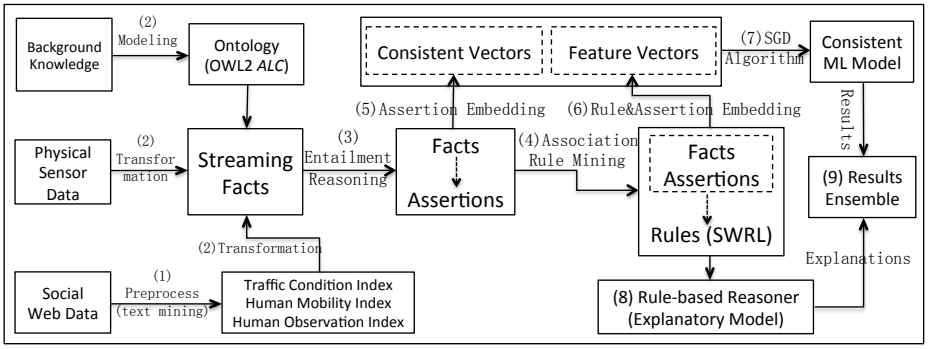

Fig. 1. High-level technical framework.

Consistent Vector. A consistent vector $V_c^i = (V_c^i(1), V_c^i(2), ..., V_c^i(d_c))$ represents the true or false value of d_c classification assertions in i^{th} snapshot which is also known as a training example. The element $V_c^i(k)$ is assigned to 1 if k^{th} assertion is positive (e.g., $GoodAir(a)$) in that snapshot, and to 0 if k^{th} assertion is negative (e.g., $\neg Cloudy(m)$). We transform the consistent vector into autocorrelation weight of the training example by counting the equal elements, and then incorporate the weight with the model using weighted SGD algorithm. The built consistent model has been proven to solve the concept shift problem in supervised learning and achieve higher accuracy than pure ML models.

Feature Vector. A feature vector $V_f^i = (V_f^i(1), V_f^i(2), ..., V_f^i(d_f))$ represents the real value of d_f attribute facts or classification assertions in i^{th} snapshot. An attribute fact (e.g., $hasAQIValue(a, 70)$) produces a real value feature, while a classification assertion (e.g., $Cloudy(m)$) generates a discrete value feature. A technique called rule embedding is developed for automatic feature extraction. The rule with high confidence and support, also known as an prominent rule indicates strong predictive information and its prefixes are used to infer effective features. For example, if the rule $Emergent(a_{t+12}) \leftarrow Hazardous(a_t) \wedge Cloudy(m_t)$ is prominent, the attribute facts $hasCloudValue$ and $hasAQIValue$ are used for real value features, and a new concept combing $Cloudy$ and $Hazardous$ is constructed for a discrete value feature.

4 Acknowledgement

This work is funded by NSFC 61473260, national key S&T Special projects 2015ZX03003012, and supported by the Fundamental Research Funds for the Central Universities.

References

1. Chen, J., Chen, H., Hu, D., Pan, J.Z., Zhou, Y.: Smog disaster forecasting using social web data and physical sensor data. In: Big Data (Big Data), 2015 IEEE International Conference on. pp. 991–998. IEEE (2015)
2. Djuric, N., Kansakar, L., Vucetic, S.: Semi-supervised learning for integration of aerosol predictions from multiple satellite instruments. In: IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013 (2013)
3. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the 22nd international conference on World Wide Web. pp. 413–422. International World Wide Web Conferences Steering Committee (2013)
4. Konkel, L.: The view from Afar: Satellite-derived estimates of global PM2.5. Environmental health perspectives 123(2), A43 (2015)
5. Lécué, F., Pan, J.Z.: Predicting knowledge in an ontology stream. In: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. pp. 2662–2669. AAAI Press (2013)
6. Lécué, F., Pan, J.Z.: Consistent knowledge discovery from evolving ontologies. In: Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 189–195 (2015)
7. Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M., Tommasi, P.: Smart traffic analytics in the semantic web with STAR-CITY: scenarios, system and lessons learned in Dublin City. Web Semantics: Science, Services and Agents on the World Wide Web 27, 26–33 (2014)
8. Pope III, C.A., Dockery, D.W.: Health effects of fine particulate air pollution: lines that connect. Journal of the air & waste management association 56(6), 709–742 (2006)
9. Ren, Y., Pan, J.Z.: Optimising ontology stream reasoning with truth maintenance system. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 831–836. ACM (2011)
10. Zheng, Y., Liu, F., Hsieh, H.P.: U-Air: When urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1436–1444. ACM (2013)
11. Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., Li, T.: Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2267–2276. ACM (2015)