Discovering air quality patterns in urban environments

Discovering Air Quality Patterns in Urban Environments

Luca Cagliero, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, Giuseppe Ricupero

Dipartimento di Automatica e Informatica,
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
E-mail: *luca.cagliero@polito.it*

Keywords Smart Cities, Data Mining, Pollutant Data, Sensor Networks.

Abstract *Monitoring air quality is currently a critical issues in smart cities. Air pollution-related data are commonly acquired through sensors deployed throughout the city area. To analyze these data collections, data analytics algorithms should be combined with reporting tools for discovering critical conditions and informing citizens and municipality actors. This paper proposes a new data mining engine to discover air quality patterns from air pollution-related data. This class of patterns includes many established patterns proposed in the data mining literature. In this study, we focused on a specific type of patterns, namely the frequent weighted itemsets, to identify combinations of pollutants that are, on average, in a critical condition. To show the usefulness of the proposed approach, the proposed engine was tested on real data acquired in a major Italian city.*

1 Introduction

Nowadays Smart Cities are increasingly pervaded by sensors deployed in public areas, on vehicles, and on wearable devices. These sensor networks allow us to collect a variety of data useful for monitoring the factors influencing the quality of citizen's life from different viewpoints. Counteracting the presence of high levels of pollutants is crucial for guaranteeing the livability of urban environments, especially because vehicular traffic, heating systems, and rejects of industrial productions significantly contribute to pollutant emissions and, thus, they have an impact on public health. The quality of the air can vary over time and across different areas of the same city. Analyzing the air quality levels acquired by sensors is particularly useful for characterizing pollutant concentrations (see Figure 1). Sensor data are usually sampled at fairly high frequencies, for relatively long time periods, and across potentially large city areas. Several data mining solutions have been proposed to analyze pollutant data in the urban environment. For example, classification techniques have been applied to perform predictions of future air quality levels Zheng et al. [2015] or predictions in city areas not equipped with monitoring stations Zheng et al. [2013], while the correlations between pollutants and meteorological data have been studied in Elminir [2005].

Main monitored pollutants	
Particulate Matter 10µm	PM_{10}
Particulate Matter 2.5µm	$PM_{2.5}$
Ozone	O_3
Nitrogen Dioxide	NO_2
Carbon Monoxide	CO
Benzene	C_6H_6

Figure 1: Monitoring air pollution.

This paper presents AiR QUAlity patTern Analyzer (ARQUATA), a data mining based engine aimed at characterizing the concentrations of air pollutants through the analysis of the sensor measurements and reporting critical conditions to citizens and municipality actors. Unlike previous approaches (e.g. Elminir [2005], Zheng et al. [2013, 2015]) it proposes (i) The use of a class of data mining patterns, hereafter denoted as *air quality patterns*, to discover combinations of pollutants whose concentration levels are averagely critical in a given spatio-temporal context (e.g. the sensor measurements acquired in a given city area during the last year). Among the patterns available in data mining literature, in this study we considered the frequent weighted itemsets, because, unlike traditional pattern mining approaches, weighted itemset mining algorithms inherently handle numerical pollutant levels. This simplifies the preprocessing phase of the analyzed data and reduces the bias due to discretization. (ii) The generation of automatic reports, which indicate the presence of critical conditions in specific contexts and their temporal evolution, by performing a comparison between the results of different mining sessions scheduled in consecutive time periods or in different city areas. The proposed engine was validated on real data collected in a major Italian city to demonstrate the usefulness of the proposed approach in a real Smart City context.

2 The ARQUATA engine

Figure 2 summarizes the main blocks of the proposed engine, which relies on three main components: (i) Data integration and preparation. (ii) Air quality pattern mining. (iii) Reporting.

Data integration and preparation. Different geo-referenced sensor networks are exploited to periodically monitor the concentration levels of the main air pollutants in the urban environment. Since each network may adopt a different timeline in sampling pollutant concentrations in different city areas, the acquired measures are then integrated by considering a common time granularity. For the sake of simplicity, hereafter we will consider daily time granularities. Each pollutant level is characterized by (i) a sampling timestamp and (ii) a set of geo-coordinates of acquisition (i.e., latitude and longitude). Pollutant concentrations are first cleaned to

remove missing values and incorrect readings, normalized to make the pollutant concentrations distributions uniform with each other, and then integrated into different contextual datasets, one for each context under analysis. A *contextual dataset* collects sensor measurements acquired from a subset of sensors corresponding to a specific spatial or functional domain (e.g. the sensors belonging to the same district, industrial area, or residential zone). Each contextual dataset consists of a set of records, one for each sampling timestamp. Each record comprises all the pollutant concentrations acquired at the corresponding timestamp.

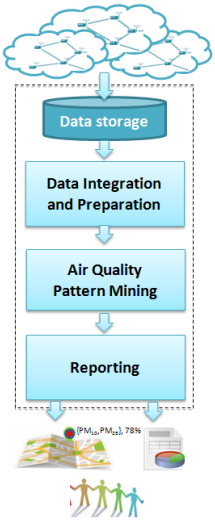

Figure 2: The ARQUATA engine.

Air quality pattern mining. This step focuses on extracting patterns characterizing the air quality. Air quality pattern extraction relies on itemset mining, an established data mining technique to discover significant yet hidden correlations among large datasets Agrawal et al. [1993]. Using itemset mining, several types of air quality patterns can be extracted (e.g. closed itemsets, generalized itemsets, emerging patterns). In this study, we focus on a specific type of itemset, namely the frequent weighted itemsets, to discover combinations of pollutants whose concentration levels are all averagely critical in the considered time period. For each pollutant a reference *critical level* is given by the domain expert (e.g., the critical level specified by law). *Weighted itemsets* are sets of items, each one representing a distinct pollutant (e.g., {$PM_{2.5}$, PM_{10}}). Weighted itemsets are extracted from contextual datasets. To this aim, for each record a weight is associated with each pollutant occurring in the itemset. Weights are computed as the percentage variation of the corresponding pollutant concentration with respect to the critical level. For example, if the critical level of $PM_{2.5}$ is 10 and the $PM_{2.5}$ level at the considered timestamp is 15 then the item weight is 50%, because the level exceeds the critical level by 50%. Weighted itemsets are characterized by a notable quality measure, denoted *critical gap*. It indicates the average percentage variation of the least pollutant weight (i.e., the minimal percentage variation of the corresponding pollutants). Using the algorithm recently proposed in Cagliero and Garza [2014], we extract from each contextual dataset all the frequent weighted itemsets, which are combinations of pollutants whose critical gap is above a given (user-specified) threshold. For example, if the critical gap threshold is set to 10%, itemset {$PM_{2.5}$, PM_{10}} is extracted if both $PM_{2.5}$ and PM_{10} have an average percentage variation above 10%.
Reporting. Since the correlations among pollutant levels can vary over time and space, users are commonly interested in monitoring their temporal and spatial evolution. To effectively characterize the underlying trends in air pollution-related data, the results of different mining sessions on data acquired from the same context are compared with each other and reported to citizens and municipality actors. Technical reports directed to domain experts include (i) Periodic summaries on the latest mining results (e.g., the top-10 combinations of pollutants in order of decreasing critical gap). (ii) A comparison between the mining results achieved in different time periods or in different city areas (e.g. the combinations of pollutants in common for all years/areas, the combinations appearing in just one year/area). Based on the above results, experts may process the extracted patterns and schedule the periodic forwarding of higher-level reports, directed to either citizens or municipality actors, discussing (i) the currently critical levels of pollutants for each city area/district (e.g., in the last winter the $PM_{2.5}$, PM_{10} concentrations were averagely critical at the same time with a critical gap higher than 30%), (ii) the most significant temporal trends in pollutant level variations (e.g., the criticality of the levels of $PM_{2.5}$, PM_{10} is constantly decreasing from year 2014 on), (iii) the most significant spatial trends (e.g., the criticality of the levels of $PM_{2.5}$, PM_{10} is more significant in the city center). Reports are tailored to end user roles (e.g. city major, assessor, citizen) and the corresponding authorities.

Autumn

itemset	*Critical gap (%)*
{$PM_{2.5}$}	90.62
{PM_{10}}	88.09
{PM_{10},$PM_{2.5}$}	77.64
{NO_2}	54.64
{NO_2,PM_{10}}	38.37
{NO_2,$PM_{2.5}$}	35.51
{NO_2,$PM_{2.5}$,PM_{10}}	34.12

Winter

itemset	*Critical gap (%)*
{$PM_{2.5}$}	120.72
{PM_{10}}	99.49
{PM_{10},$PM_{2.5}$}	93.49
{NO_2}	76.87
{NO_2,$PM_{2.5}$}	53.36
{NO_2,PM_{10}}	50.51
{NO_2,$PM_{2.5}$,PM_{10}}	45.92

Spring

itemset	*Critical gap (%)*
{NO2}	30.07

Summer

itemset	*Critical gap (%)*
no patterns satisfying the threshold	

Table 1: Year 2013. Seasonal analysis: Frequent weighted itemsets. Critical gap thr.=30%.

3 Experimental results and discussion

The proposed approach was validated on real pollutant data acquired on a daily basis by the ARPA Lombardia in the central area of Milan (Italy). We analyzed the seasonal trends in pollutant data acquired in year 2013. For each pollutant we considered as reference critical level the highest safe concentration (i.e., the upper border of the green zone available on the ARPA website). To characterize seasonal trends in pollutant data, we extracted all the weighted patterns whose critical gap is higher than 30% (see Table 1). Autumn and winter are seasons characterized by similar trends. In both seasons the levels of $PM_{2.5}$ and PM_{10} are significantly higher than the corresponding critical levels (e.g., the level of $PM_{2.5}$ in autumn and winter are on average 90.62% and 120.72% higher than the critical level, respectively). The combination of pollutants that simultaneously exceed the critical level in most cases are particularly interesting. For instance, based on the critical gap of pattern {PM_{10},$PM_{2.5}$}, in the winter season pollutants PM_{10} and $PM_{2.5}$ simultaneously exceed their critical level by 93.49%. Since the two critical conditions appeared to be strongly correlated with each other, targeted actions may be performed to counteract them. A fair correlation between the triple of pollutants NO_2, PM_{10}, and $PM_{2.5}$ appeared as well. However, the critical gap of patterns {NO_2,PM_{10}} (53.56%) and {NO_2,$PM_{2.5}$} (50.51%) are significantly lower than those of {PM_{10},$PM_{2.5}$} (93.49%). The above results are consistent with the expectation, because the association between PM_{10} and $PM_{2.5}$ is established. Note that pattern {PM_{10},$PM_{2.5}$} is extracted only in autumn and winter, because the pollutant concentrations are probably related to the use of heating systems. The municipality may foster the use of new generation heating systems, characterized by lower pollutant emissions through targeted actions and then use the framework to analyze the effect of the performed actions on the air quality. For example, a yearly comparison between the critical gaps of patterns {PM_{10},$PM_{2.5}$}, {NO_2,$PM_{2.5}$}, and {NO_2,PM_{10}} among years 2013, 2014 and 2015 is reported in Figure 3. The comparison showed a slight decrease from 2013 to 2014, but a slight increase from 2014 to 2015.

4 Conclusions

In this paper we proposed the preliminary version of the ARQUATA system, which mines air quality patterns from air pollution-related data and reports critical conditions to citizens and municipality actors. As future work, we plan to (1) define novel types of air quality patterns based on other itemset mining algorithms (e.g. the closed itemsets) and (2) study the impact of people's mobility and public transports on the air quality.

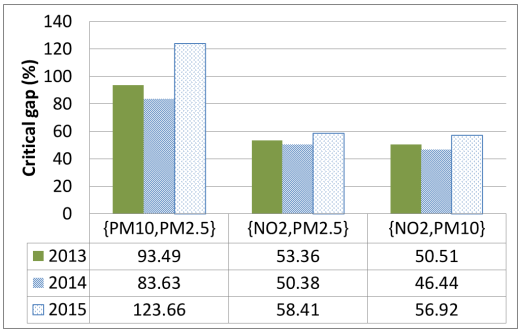

Figure 3: Comparison between winters 2013, 2014, and 2015.

5 Acknowledgements

The research leading to these results has received funding from the Italian Ministry of Research - cluster *Tecnologie per le Smart Communities, Progetto MIE - Mobilità Intelligente Ecosostenibile.*

References

R. Agrawal, T. Imielinski, and Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD 1993*, pages 207–216, 1993.

L. Cagliero and P. Garza. Infrequent weighted itemset mining using frequent pattern growth. *IEEE Trans. on Knowledge and Data Engineering*, 26(4):903–915, 2014. ISSN 1041-4347.

Hamdy K. Elminir. Dependence of urban air pollutants on meteorology. *Science of The Total Environment*, 350(1-3):225 – 237, 2005. ISSN 0048-9697.

Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: when urban air quality inference meets big data. In *ACM SIGKDD*, pages 1436–1444, 2013.

Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *ACM SIGKDD*, pages 2267–2276, 2015.