Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2018) 26: 1302 – 1313

doi:10.3906/elk-1707-99

Improvement of air pollution prediction in a smart city and its correlation with weather conditions using metrological big data

Talat ZAREE, Ali Reza HONARVAR*
Department of Electrical and Computer Engineering, Safashahr Branch, Islamic Azad University, Safashahr, Iran

Received: 08.07.2017 • **Accepted/Published Online:** 14.11.2017 • **Final Version:** 30.05.2018

Abstract: Smart cities are an important concept for urban development. This concept addresses many current critical urban problems including traffic and environmental pollution. As utilization of the Internet of things and technology in smart cities increases, large volumes of big data are generated and collected by sensors embedded at different places in the city, which present a real-time display of what is happening throughout the city at all times. Such data should be processed and analyzed as a response to ensure effectiveness and improvement in quality of provided services; correct use and analysis of such data is valuable. Big data mining is the most effective method for analyzing such data. In this paper, aiming to increase speed and accuracy in predicting real levels of air pollution, its location, and effects of weather conditions on density of air pollution, a K-means clustering algorithm using the Mahout library is used as a big data mining tool on datasets of a city pulse project. Results of this study show that temperature, low air pressure, relative increase in moisture, and wind speed are causes of low pollution density at the cleanest point of the city. The SSE evaluation metric shows the high speed of this clustering method for big data, and results obtained from employing RMSE = 0.632 and MSE = 0.488 statistical measures indicate the high efficiency and accuracy of the proposed method in predicting air pollution.

Key words: Smart city, Internet of things, big data mining, K-means clustering, Mahout, air pollution

1. Introduction

Today, smart cities are admired as promising solutions to future challenges for providing better services to all citizens and improving efficiency. According to some researchers, the concept of smart cities is a result of a smart growth movement in urban management policies in the late 1990s with the support of urban planners in Portland, USA [1]. Although effects of smart cities have spread throughout the world, the definition is still ambiguous; in other words, no common definition has been proposed for a smart city and describing a global standard definition is difficult. However, most significant definitions specify common characteristics, features, and components of smart cities. Some definitions of a smart city are as follows:

- A city in which investment in human and social resources, conventional (transportation) and modern (ICT) fuel communication infrastructures, sustainable economic growth, and high quality of life is conducted through wise management of natural resources by participatory government is said to be smart [2].
- A smart city is a city in which there are 6 main components including smart economy, smart transportation, smart environment, smart citizens, smart life, and smart management [3].

*Correspondence: alireza_honarvar@yahoo.co.uk

- A smart city should employ smart computation technology such that important components, infrastructures, and services of a city are more smart and more efficient [4].

A common point of combinational definitions of information and communication technology (ICT) with investment in human and social resources, services, and modern urban infrastructures is for creating sustainable economic growth and high quality of life for citizens. To this end, air pollution is one of the most important factors that affect quality of life in urban areas and it threatens human health significantly, especially for vulnerable strata (children, the elderly, and patients). Weather conditions affect air pollution significantly. Specific weather conditions create and escalate critical air pollution conditions. Identifying the relation between weather conditions and air quality provides the possibility for researchers to minimize bad effects of air pollution.

All smart cities have a smart system architecture of integrated sensors, software, and networks, which is called the Internet of things (IoT) [5]. Smart objects can communicate with each other through the Internet based on new ICT for collecting and exchanging information. With the ever increasing growth of the IoT and technology in smart cities, a large volume of data is generated and collected in different formats through sensors installed at different points. These sensors are used for monitoring city status and collecting data about status of cities, which present a real-time representation of what happens throughout the city. Analysis and mining of data from dynamic cities is an inevitable step towards constructing a smart city [6]. Such data is a good example of big data [7].

"Big data" was first proposed in 1998 in "Silicon Graphic" by John Mashey as "big data and the next wave of infrastructure stress". Later, in 2000, the first scientific paper in which "big data" was mentioned was published by Diebold. The origin of big data returns to the fact that a large volume of data is generated every day. At the beginning of 2001, Dug Lung, analyzer of Gartner Institute, introduced three features including volume, velocity, and variety for defining big data in a report about the challenges and opportunities of increasing data. Gartner and many other institutes still use these features to define big data. In 2012, Gartner updated its definition of big data as follows: "big data is a large volume, high velocity, high acceleration and high variety of information which requires a new form of processing to enrich decisions, explore a new vision and optimize procedures". Today, new concepts like validity, variability, and value are also proposed for big data [8].

Correct use and effective analysis of data is valuable for urban planning and decision making, which improves quality of life through improving efficiency of services. Knowledge of using such data is called data mining. Data mining is the procedure of exploring interesting patterns and knowledge among a large volume of data. In other words, data mining explores a "model" for data [9], which is used in a wide variety of applications like business, medicine, science, and engineering, which results in better services in many business contexts for service providers and consumers. Since software infrastructures for smart cities should provide computational capabilities with high efficiency that are scalable and reliable and big data are mainly stored in different locations and the volume of stored data is growing continuously, a distributed computation platform is required [10].

From the data mining viewpoint, even if big data are of great hidden value, exploring knowledge from big data is very challenging because knowledge exploration procedures and data mining are designed for usual datasets. The negative aspect of current data mining techniques used for big data is insufficient and parallel scalability. In general, available data mining techniques face large problems, which necessitates handling unprecedented heterogeneity, volume, velocity, privacy, accuracy, and future trust along with big data and big data mining. Improving available techniques using parallel processing architecture and distributed storage systems will overcome the aforementioned challenges and will change future of data mining technology [11].

One of the big data mining tools is Mahout Apache [12]. The Mahout subproject, which is an inseparable component of the Hadoop project, is an environment for creating distributed machine learning programs and algorithms. This subproject includes different algorithms and libraries for data mining, among which the K-means clustering algorithm can be mentioned. The K-means clustering algorithm is a known clustering algorithm used in a wide variety of problems limited to small datasets. Currently, many of the datasets are large and cannot be stored or processed in the main cache. The Hadoop platform and library of optimal Mahout algorithms is a cheap and promising solution for improving conventional clustering algorithms and solving problems of large datasets [13].

In this paper, in order to increase speed and accuracy in predicting density of air pollution and its relation with weather conditions and finding the cleanest and most polluted areas in a smart city, the K-means clustering algorithm of the Mahout library is used. The rest of this paper is organized as follows: Section 2 reviews previous works in the context of predicting air pollution using data mining algorithms. Section 3 presents the proposed method. Section 4 analyzes the method and Section 5 concludes the paper and presents some suggestions for future works.

2. Related works

Previous studies in the context of predicting air pollution using data mining have mainly used conventional data mining algorithms, which do not perform well while facing big data. The authors of [14] employed an artificial neural network and a backpropagation algorithm to investigate air pollution at two points of a city. Meteorology data like temperature, wind speed, and relative moisture were considered as input parameters; densities of nitrogen dioxide, sulfur dioxide, and inhaling suspended particles were considered as output parameters. Evaluation results using minimum square error showed that moisture was the most effective parameter in air pollution at both points. The authors of [15] evaluated air pollution levels at different points of a city using longitude and latitude by applying the K-means clustering algorithm and selecting numbers of iterations for k from 3 to 10 for air pollution data. The authors of [16] used neural network and PCA techniques as a useful tool for modeling air pollution prediction. Eight parameters of air pollution quality were collected at 10 monitoring stations in Malaysia for 7 years (2005–2011) for identifying pollution sources, in which PCA was used to reduce data dimensions from 8 to 5 important parameters including O3, PM10, CH4, NMHC, and THC. By applying a neural network, it was concluded that prediction is performed better with fewer variables and the MSE of 10.017 was obtained. The authors of [17] showed that rain and temperature in different seasons were associated with air pollutants by studying air pollution through sequential clustering of an air pollution daily index from 2004 to 2007 and it was found that reducing wind speed in areas like the Yangtze River delta is one of the factors that increase the density of suspended particles.

3. Proposed method

3.1. Datasets

This paper is based on air pollution data and weather data obtained from a big data analysis system in a smart city [18] called the CityPulse open dataset, which belongs to Brasov in Romania. Air pollution data are collected by air pollution sensors installed beside traffic sensors at 449 points of the city and include longitude and latitude of that position along with 5 main components of air pollution and information regarding Brasov's weather including 4 categories of meteorology data. Combining these two datasets results in a new dataset including amount of ozone, particulate matter, carbon monoxide, sulfur dioxide, nitrogen dioxide, longitude,

latitude, timestamp, humidity, temperature, wind speed, and air pressure. Table 1 presents one of these files as an example.

3.2. K-means algorithm of Mahout Apache

This algorithm is one of the most important and applicable clustering algorithms that has high time complexity while facing big datasets and limited complexity on machine memory. For this reason, in big datasets, Hadoop and Map/Reduce platforms are used to increase operation speed. Implementation of this algorithm requires converting input data of HDFS, which is CSV, into a vector. The K-means algorithm has the following input parameters:

- a SequenceFile including input vector
- a SequenceFile including centers of initial cluster
- a metric for measuring similarity
- convergence threshold
- maximum number of processing iterations

The output of the algorithms is also a SequenceFile.

There are three stages in a K-means job:

Initial stage: segmentation of the dataset into HDFS blocks; their replication and transferring to other machines; according to the number of blocks and configuration of the cluster, necessary tasks will be assigned to it.

Map stage: calculates the distances between samples and centroids; matches samples with the nearest centroid and assigns them to that specific cluster.

Reduce stage: recalculates the centroid point using the average of coordinates of all points in that cluster. Averages of the associated points are used to produce new locations of the centroid. The configuration of the centroids is fed back into the Mappers. The loop ends when the centroids converge [13].

3.3. Implementing the proposed method

After scanning the data and making them ready using the Hadoop platform and K-means clustering algorithm, data are clustered. This algorithm is performed in the following steps:

Step 1: Making data ready and converting them into CSV format.

Step 2: Converting data into SequenceFile vector.

Step 3: Executing Mahout K-means code and specifying number of clusters from 3 to 9.

Step 4: Preparing output of clusters after executing k-means algorithm code and considering clustering. Output results are presented in Tables 2–8. In order to find low and high pollution areas in clusters, focus is on analyzing ozone and its density in those areas. Ozone, which is the main component of smoky fog, is a gas created through the combination of nitrogen oxide and hydrocarbons in the presence of sunlight. Lung damage caused by ozone pollution threatens every three people out of five. Most people do not know that smoky fog threatens not only humans but also other creatures. Ozone smoky fog damages trees and agricultural products in many areas.

Step 5: Analysis and evaluation of outputs. A clustering evaluation metric is used to specify optimal clustering.

Table 1. An example of a dataset.

ID	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Longitude	Latitude	Timestamp	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
1	101	94	49	44	87	10.1049	56.2317	2014/8/1 12:05 Am	68	18	7.4	1012
2	106	97	48	47	86	10.1049	56.2317	2014/8/1 12:10 Am	68	18	7.	1012
...	....	...	...	....	...	...	...	...	...	...	...	...
17568	60	146	55	199	127	10.1049	56.2317	2014/10/1 11:55 Am	77	14	14.8	1025
17569	61	145	53	204	126	10.1049	56.2317	2014/10/1 12:00 Am	77	14	18.8	1025

Table 2. Clustering results using K-means algorithm, $k = 3$.

Cluster	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
Cluster1	166	201	118	90	173	64	15	18.5	1008
	93	61	117	20	79	68	18	5.6	1012
	70	21	48	180	180	77	13	7.4	1022
	147	135	61	193	156	62	17	5.5	1014
	36	98	89	183	90	77	16	14.8	1013
Cluster2	189	200	163	110	138	45	20	4	1008
	172	107	77	106	120	80	12	3.7	1017
Cluster3	204	50	24	125	28	35	23	11.1	1.14
	162	82	164	197	155	73	18	11.1	1017

Table 3. Clustering results using K-means algorithm, $k = 4$.

Cluster	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
Cluster1	204	50	24	125	28	35	23	11.1	1.14
	162	82	164	197	155	73	18	11.1	1017
Cluster2	166	201	118	90	173	64	15	18.5	1008
	93	61	117	20	79	68	18	5.6	1012
Cluster3	189	200	163	110	138	45	20	4	1008
	172	107	77	106	120	80	12	3.7	1017
Cluster4	70	21	48	180	180	77	13	7.4	1022
	147	135	61	193	156	62	17	5.5	1014
	36	98	89	183	90	77	16	14.8	1013

Table 4. Clustering results using K-means algorithm, $k = 5$.

Cluster	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
Cluster1	204	50	24	125	28	35	23	11.1	1.14
	162	82	164	197	155	73	18	11.1	1017
Cluster2	166	201	118	90	173	64	15	18.5	1008
	93	61	117	20	79	68	18	5.6	1012
Cluster3	70	21	48	180	180	77	13	7.4	1022
	147	135	61	193	156	62	17	5.5	1014
	36	98	89	183	90	77	16	14.8	1013
Cluster4	89	206	116	145	96	68	19	14.8	1014
Cluster5	189	200	163	110	138	45	20	4	1008
	172	107	77	106	120	80	12	3.7	1017

Table 5. Clustering results using K-means algorithm, $k = 6$.

Cluster	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
Cluster1	166	201	118	90	173	64	15	18.5	1008
	93	61	117	20	79	68	18	5.6	1012
	70	21	48	180	180	77	13	7.4	1022
	147	135	61	193	156	62	17	5.5	1014
	36	98	89	183	90	77	16	14.8	1013
Cluster2	130	132	187	36	134	77	15	9.3	1007
Cluster3	134	198	61	23	92	88	12	1.9	1024
Cluster4	204	50	24	125	28	35	23	11.1	1.14
	162	82	164	197	155	73	18	11.1	1017
Cluster5	189	200	163	110	138	45	20	4	1008
	172	107	77	106	120	80	12	3.7	1017
Cluster6	121	117	107	198	58	49	18	13	1008

Table 6. Clustering results using K-means algorithm, $k = 7$.

Cluster	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
Cluster1	204	50	24	125	28	35	23	11.1	1.14
	162	82	164	197	155	73	18	11.1	1017
Cluster2	166	201	118	90	173	64	15	18.5	1008
	93	61	117	20	79	68	18	5.6	1012
Cluster3	86	23	139	171	134	69	11	14.8	1010
Cluster4	36	98	89	183	90	77	16	14.8	1013
Cluster5	189	200	163	110	138	45	20	4	1008
	172	107	77	106	120	80	12	3.7	1017
Cluster6	89	206	116	145	96	68	19	14.8	1014
Cluster7	70	21	48	180	180	77	13	7.4	1022

Table 7. Clustering results using K-means algorithm, $k = 8$.

Cluster	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
Cluster1	189	200	163	110	138	45	20	4	1008
	172	107	77	106	120	80	12	3.7	1017
	166	201	118	90	173	64	15	18.5	1008
	93	61	117	20	79	68	18	5.6	1012
	204	50	24	125	28	35	23	11.1	1.14
Cluster2	87	29	157	120	146	72	13	11.1	1009
Cluster3	70	21	48	180	180	77	13	7.4	1022
Cluster4	36	98	89	183	90	77	16	14.8	1013
Cluster5	89	206	116	145	96	68	19	14.8	1014
Cluster6	121	117	107	198	58	49	18	13	1008
Cluster7	130	132	187	6	134	77	15	9.3	1007
Cluster8	134	198	61	23	92	88	12	1.9	1024

Table 8. Clustering results using K-means algorithm, $k = 9$.

Cluster	Ozone	Particulate_matter	Carbon_monoxide	Sulfur_dioxide	Nitrogen_dioxide	Hum (%)	Temp (°C)	Wspdm (km/h)	Pressure (mbar)
Cluster1	189	200	163	110	138	45	20	4	1008
	172	107	77	106	120	80	12	3.7	1017
	166	201	118	90	173	64	15	18.5	1008
	93	61	117	20	79	68	18	5.6	1012
	36	98	89	183	90	77	16	14.8	1013
Cluster2	204	50	24	125	28	35	23	11.1	1014
	162	82	164	197	155	73	18	11.1	1017
Cluster3	147	135	61	193	156	72	13	5.5	1014
Cluster4	87	29	157	120	146	72	13	11.1	1009
Cluster5	134	198	61	23	92	88	12	1.9	1024
Cluster6	70	21	48	180	180	77	13	7.4	1022
Cluster7	130	132	187	6	134	77	15	9.3	1024
Cluster8	99	93	103	84	58	72	15	9.3	1017
Cluster9	89	206	116	145	96	68	19	14.8	1014

4. Experiments and results

4.1. Evaluation metric

As observed in the results of clusters, the minimum amount of pollution is 36, which is minimum in all clusters from $k = 3$ to $k = 9$ and only placement priority in top terms of the clusters is different. In order to specify the optimal cluster, a clustering evaluation metric is used. One of the important problems in clustering is to decide about the best cluster set for a dataset in terms of number of clusters and membership in clusters. For this purpose, Eq. (1) uses the sum of square errors (SSE) method, which is a measure of quality inside the cluster. The lower the SSE is, the higher intercluster similarity is and the better the result is.

$$\mathbf{SSE} = \sum_{\mathbf{k=1}}^{\mathbf{k}} \sum_{\mathbf{i=1}}^{|\mathbf{c_k}|} (\mathbf{Dist}(\mu_{\mathbf{k}}, \mathbf{x_i}))^{\mathbf{2}} \tag{1}$$

$\mathbf{C_k}$ = size of cluster - number of samples $\mathbf{U_k}$ = center of cluster

In clustering contexts, one is always about to reduce clusters for better planning and control. In addition, in clustering big data, a cluster that has more top terms is better compared to other clusters. For this reason, by investigating values of the evaluation metric in each cluster and specifying the minimum value of the metric at any number of iterations and comparing them, the minimum value is obtained at $k = 3$ and cluster 1 with evaluation metric of 25.4. Figure 1 compares evaluation metrics in clusters.

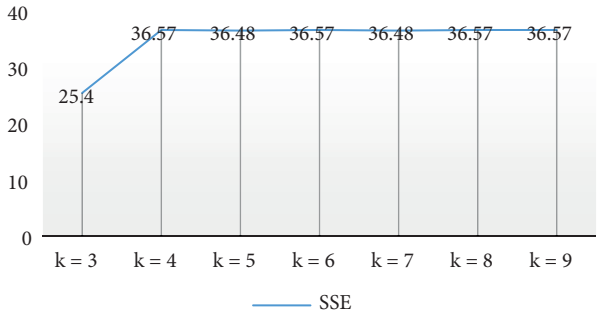

Figure 1. Comparing SSE evaluation metric in clusters.

As can be seen, a high number of iterations is not required and at $k = 3$ and the minimum number of iterations, the evaluation metric is minimum and continuing clustering is not necessary and clustering is performed fast. Air pollution density in the most appropriate cluster is 36. This amount of pollution is associated with the cleanest area of the city with longitude of 10.22863922 and latitude of 56.1976518 with 77% moisture and 16 °C temperature where wind speed is 14.8 and pressure is 1013. Maximum pollution density is 204, which has appeared in all clusters and belongs to the most polluted area of the city with longitude of 10.22311078 and latitude of 56.19038745 with 35% moisture and 23 °C temperature where wind speed is 11.1 and pressure is 1014.

By investigating weather conditions in the most polluted area and cleanest area of the city, it has been concluded that relative moisture in the cleanest area is higher while temperature and pressure in the clean area is lower than in the polluted area. Wind speed is another factor that affects the density of pollutants and results show that high wind speed in the clean area reduces density of pollutants.

After finding the optimal clustering, performance and efficiency of the proposed method based on RMSE and MSE are evaluated according to following equations.

$$\mathbf{MSE}=\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{f_i}-\mathbf{y_i})^{\mathbf{2}} \tag{2}$$

$$\mathbf{RMSE}=\sqrt{\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{f_i}-\mathbf{y_i})^{\mathbf{2}}} \tag{3}$$

RMSE indicates the deviation of predicted values from observed values and the lower bound of RMSE is zero. MSE also indicates accuracy of the model, which can vary from zero to infinity. The closer these values are to zero, accuracy is higher. With RMSE = 0.632 and MSE = 0.488, the proposed method not only obtains considerable results but also obtains high accuracy and efficiency.

5. Discussion

In this paper, aiming to increase the speed and accuracy of predicting air pollutants and improving air pollution management in a smart city, the K-means clustering algorithm of the Mahout library is proposed.

Considering studies performed in the context of predicting air pollution, conventional clustering algorithms and techniques are not designed for smart cities and are not suitable in this context. Accordingly, improving techniques and algorithms through distributed computation architecture is necessary. In order to investigate this issue through averaging the air pollution dataset, the K-means algorithm is implemented using Rapid Miner software and iterations are continued to $k = 9$ for finding suitable clustering where the evaluation metric for this clustering is shown in Figure 2, which shows suitable clustering for $k = 9$.

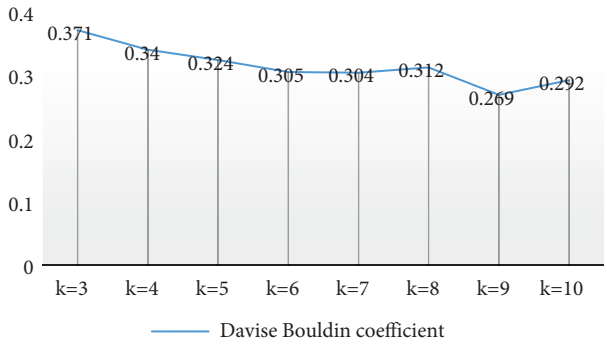

Figure 2. Evaluating clustering with Davis coefficient.

While facing big data, accuracy is lowered because all data cannot be employed in conventional clustering. Minimum value of the Davies index is 0.28, which shows $k = 9$ and the cleanest area with longitude of 10.17142 and latitude of 56.15884 with density of 89 is located in this cluster. The authors of [15] determined optimal clustering with $k = 10$ and air pollution of 89 using conventional clustering. In the proposed method, speed and accuracy increase as shown in the previous section through plotting the evaluation index of clustering, in which a high number of iterations is not required and a minimum evaluation index of $k = 3$ is shown unlike conventional data mining, and least pollution density is also more accurate and equal to 36. This comparison is shown in Figure 3.

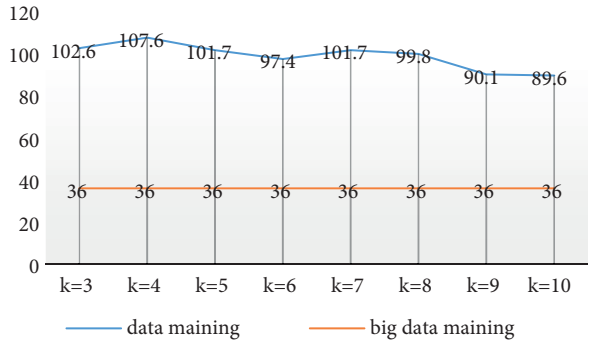

Figure 3. Comparison of clustering in data mining and big data mining (proposed method).

6. Conclusion

In this study, in order to predict air pollution and correlation of its density with weather conditions and improve air pollution management in a smart city, the K-means clustering algorithm of the Mahout library is employed for air pollution data and the CityPulse project is used for weather conditions. Number of iterations for k is considered to vary from 3 to 9 and analysis of clusters is performed based on ozone density in the cluster. Using evaluation metric SSE = 25.4 and finding the optimal clustering, the most polluted area with density of 204 and cleanest area with density of 36 are found at $k = 3$. In previous methods and conventional clustering, least pollution was specified at $k = 10$ with density of 89. Thus, it is necessary to mention that conventional data mining algorithms are not efficient in handling big data of smart cities and cannot handle this volume of data. Thus, using big data mining tools like algorithms of the Mahout library might be a suitable solution. Statistical indices RMSE = 0.632 and MSE = 0.4 prove the high accuracy and efficiency of the proposed method in predicting air pollution.

Investigations show that weather conditions affect air quality significantly. High pressure, temperature, and reducing relative moisture and wind speed are the factors that increase pollution density in polluted areas. Finding the most polluted and cleanest areas of the city improves citizens' quality of life, which is the main purpose of smart cities, along with which the environment is also improved significantly. Controlling air pollution is one of the main advantages of smart cities; the proposed method is applied in large cities to find polluted areas in real time and using pollution control methods in those areas helps manage the city. Air pollution is a great environmental threat for health. By decreasing air pollution, countries can reduce diseases like brain stroke, cardiac diseases, lung cancer, and chronic and acute respiratory diseases like asthma. The lower the level of pollution is, cardiac and respiratory health is better and living in these areas has lower risk and vice versa. In order to obtain more desirable results from the proposed method, using the fuzzy clustering algorithm of the Mahout library is suggested for future works.

References

[1] Harmon RR, Castro-Leon E, Bhide S. Smart cities and the Internet of things. In: IEEE Management of Engineering and Technology 2015 Portland International Conference; 2–6 August 2015; Portland, OR, USA. New York, NY, USA: IEEE. pp. 485-494.

[2] Caragliu A, Del BO CH, Nijkamp P. Smart cities in Europe. In: 3rd Central European Conference in Regional Science; 7–9 October 2009; Kosice, Slovakia. pp. 49-59.

[3] Giffinger R, Fertner CH, Kramar H, Kalasek R. Smart Cities: Ranking of European Medium-Sized Cities. Vienna, Austria: Vienna University of Technology, 2007.

[4] Wenge R, Zhang X, Dave C, Chao L, Hao SH. Smart city architecture: a technology guide for implementation and design challenges. China Communications 2014; 11: 56-69.

[5] Atzori L, Iera A, Morabito G. Internet of things survey. Computer Netw 2010; 54: 2787-2805.

[6] Pan G, Qi G, Zhang W, Li SH, Wu Z, Yang L. Trace analysis and mining for smart cities: issues, methods, and applications. IEEE Commun Mag 2013; 121: 120-126.

[7] Zikopoulos PC, Eaton C, DeRoos D, Deutsch TH, Lapis G. Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: McGraw-Hill, 2011.

[8] Fan W, Bifet A. Mining big data: current status, and forecast to the future. ACM SIGKDD Explorations Newsletter 2012; 14: 1-5.

[9] Leskovec J, Rajaraman A, Ullman J. Mining of Massive Datasets. New York, NY, USA: Cambridge University Press, 2011.

[10] Wu X, Zhu X, Wu G, Ding W. Data mining with big data. IEEE T Knowl Data En 2014; 26: 97-107.

[11] Che D, Safran M, Peng Z. From big data to big data mining: challenges, issues and opportunities. In: 18th International Conference on Database Systems for Advanced Applications; 22–25 April 2013; New York, NY, USA. pp. 1-15.

[12] Owen S, Anil R, Dunning T, Friedman E. Mahout in Action. Greenwich, CT, USA: Manning Publications, 2012.

[13] Esteves R, Pais R, Rong C. K-means clustering in the cloud - a Mahout test. In: 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications; 22–25 March 2011; Singapore. New York, NY, USA: IEEE. pp. 514-519.

[14] Vijayaraghavan N, Mohan G. Air pollution analysis for Kannur city using artificial neural network. International Journal of Science and Research 2016; 5: 1399-1401.

[15] Doreswamy, Ghoneim O, Manjaunath BR. Air pollution clustering using K-means algorithm in smart city. International Journal of Innovative Research in Computer and Communication Engineering 2015; 3: 51-57.

[16] Azid A, Juahir H, Toriman M, Amir Kamarudin M, Mohd Saudi A, Che Hasnam C, Abdul Aziz N, Azaman F, Latif M, Mohamed Zainuddin S et al. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques a case study in Malaysia. Springer Water Air Soil Pollut 2014; 225: 2063.

[17] Gao H, Chen J, Wang B, Tan SM, Lee C, Yao X, Yan H, Shi J. A study of air pollution of city clusters. Atmos Environ 2011; 45: 3069-3077.

[18] Honarvar AR, Sami A. A multi-source big data analytic system in smart city for urban planning and decision making. In: Doctoral Consortium; April 2016; Rome, Italy. pp. 32-36.