

An IOT based Novel approach to predict Air Quality Index (AQI) using Optimized Bayesian Networks

¹Krishna Chaitanya Atmakuri, ²Y Venkata Raghava Rao

¹Research Scholar, ²Professor, Department of Computer Science and Engineering

^{1,2}Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.

Email: ¹chaituit2004@gmail.com, ²yvenkataraghava1@gmail.com

<https://doi.org/10.26782/jmcms.2019.04.000035>

Abstract

As the size of the air quality data increases, it is difficult to forecast the air quality metrics due to the non-stationary and randomization form of data distribution. Air quality prediction refers to the problem of finding the air quality by using statistical inference measures. However, traditional air prediction models are based on static fixed parameters for quality prediction. Also, it is difficult to classify and predict the air quality index for both rural and urban areas due to change in data drift and distribution. $PM_{2.5}$ is one of the major factor to predict the air quality index (AQI) and its severity level. Due to high noisy and outliers in the $PM_{2.5}$ data, it is difficult to classify and predict the air quality by using the traditional quality prediction models. In order to overcome these issues, an optimized Bayesian networks based probabilistic inference model is designed and implemented on the air quality data. An IOT enabled Air pollution monitoring system includes a DSM501A Dust sensor which detects $PM_{2.5}$, $PM_{1.0}$, MQ series sensor interfaced to a Node MCU equipped with ESP32 WLAN adaptor to send the sensor reading to Thing Speak cloud. In the proposed model, the data is initially gathered from the ICAO records of Safdarjung weather station and pre-processed. An improved discrete and continuous parameter estimation and bayes score optimization are implemented on the air quality prediction process. Experimental results show that the present optimized Bayesian network classify and predicts the air quality data with high less computational error rate and high accuracy. Further the proposed optimized model is applied on the real data which is gathered using IOT enabled gas sensors and the model is giving best results in predicting the air quality Index.

Keywords : Bayesian Classification Algorithm, IOT, Air Quality Index, Data Pre-processing

I Introduction

Now-a-days, air quality plays a vital role in the air quality management system. A large number of research works have been introduced in the literature on the air

quality index prediction and air quality severity prediction. Air quality index has become a major issue in many areas of the world (fig 1) and air quality control is a highest priority to the government in air quality management system. As the number of urbanization and industrialization increases, the emission of waster gas from automobiles has become the main source of poor air quality. The presence of air quality metrics such as carbon monoxide (co), particulate matter (PM), PM2.5 (particulate matters with a diameter $\leq 2.5 \mu\text{m}$), PM10 (particulate matters with a diameter $\leq 10 \mu\text{m}$), nitrogen oxides and ozone will damage the environment, ecosystems and human health [12]. These atmospheric pollutants may lead to cardiovascular and respiratory diseases. Statistical learning models are commonly implemented in the air quality prediction, among them, auto-regressive integrated moving average models, Bayesian inference models , and multiple linear regression (MLR) models These prediction models are implemented by using statistical parametric and non-parametric methods[3].The main objective of the statistical methods is to find the essential key factors of air quality index. However, these models have limitations or issues in real time air quality data, namely these models are based on initialization parameters and they are not applicable for linear and non-linear probabilistic prediction [9].

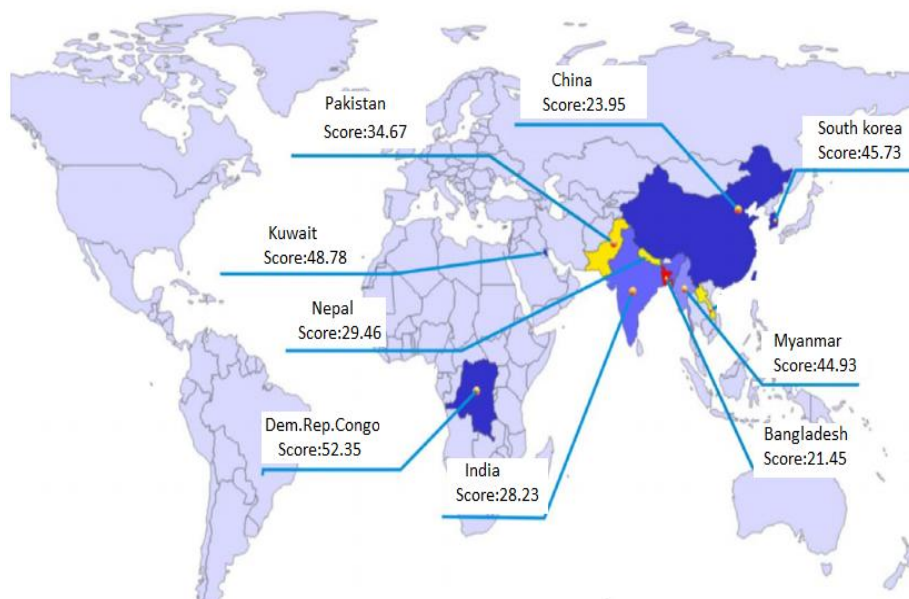


Fig. 1 worldwide Air pollution score

Poor air quality is not unique to single country; many developing countries suffer from sever air pollution as shown in fig1. Many air quality indices have been introduced in the literature to evaluate the quality of the air pollution [13]. The AQI consists of six types as shown in table 1.

S.No.	AQI Levels Of Health Concern	Numerical Value	Meaning
1.	Good	0 – 50	Air quality is considered satisfactory and air pollution poses little or no risk.
2.	Moderate	51 – 100	Air Quality is acceptable.
3.	Unhealthy for sensitive groups	101 – 150	Members of sensitive groups may experience health effects.
4.	Unhealthy	150 – 200	Every one affected, sensitive groups may experience more serious health effects.
5.	Very Unhealthy	200 – 300	Health alert everyone may experience more serious Health effects.
6.	Hazardous	300 – 500	Health warning of Emergency conditions. The entire population is more likely to be affected.

Table. 1 AQI levels and its meaning

The first index was the “pollutant standard index”(PSI) and later it was modified as air quality index(AQI) by the United States Environmental Protection Agency. Each AQI pollutants and its causes are summarized in table2.

Traditional machine learning classifiers such as SVM, naïve bayes and rule mining techniques are capable of finding essential patterns based on the air pollution severity. Statistical learning approaches such as multiple non-linear regression, log regression, support vector regression are used to estimate the air pollution index measure on the limited data [8].

Pollutants	Emission Sources		Major Effects	
	Natural Sources	Anthropogenic Sources	Health Effects	Environment Effects
Sulphur Dioxide (SO ₂)	Volcanic emissions	Burning of fossil fuels, metal melting etc.	Respiratory problems, heart and lung disorders, visual impairment	Acid rain
Nitrogen dioxide (NO ₂)	Lightning, forest fires etc.	Burning of fossil fuels, biomass & high temperature combustion process	Pulmonary disorders, increased susceptibility to respiratory infections	Precursor of ozone formation in troposphere, aerosol formation.
Particulate matter (PM)	Windblown dust, pollen spores, photochemically produced particles	Vehicular emissions, industrial combustion processes, construction industries	Respiratory problems, liver fibrosis, lung/liver cancer, heart stroke, bone problems	Visibility reduction
Carbon monoxide (CO)	Animal metabolism, forest fires, volcanic activity	Burning of carbonaceous fuels, emission from IC engines	Anoxemia leading to various cardiovascular problems. infants, pregnant women and elderly people are at higher risk.	Effects the amount of greenhouse gases which are linked to climate change and global warming.
Ozone (O ₃)	Present in stratosphere at 10-50 km height	Hydrocarbons and NO _x upon reacting with sunlight results in (O ₃) formation.	Respiratory problems, asthma, bronchitis etc.	O ₃ in upper troposphere causes green house effects, harmful effects on plants, death of plant tissues.

Table 2: AQI pollutants and its causes

The main problem in the statistical and machine learning models include; models are not applicable to predict new type of air quality metric, the prediction rate of these models are not exact and finally the error rate of the prediction models are high [8]. In order to overcome these limitations, an optimized Bayesian networks based probabilistic inference model is designed and implemented on the PM_{2.5} air quality data.

II. Related works

Regression tree (decision tree for regression) uses variances reduction metric to choose the best splitting parameter. Each separation by the chosen splitting features should give the maximal information. In math, the algorithm chooses the feature and its level/value to minimize

$$\frac{1}{|S_t|^2} \sum_{i \in S_t} \sum_{j \in S_t} \frac{1}{2} (x_i - x_j)^2 + \frac{1}{|S_f|^2} \sum_{i \in S_f} \sum_{j \in S_f} \frac{1}{2} (x_i - x_j)^2$$

Where x represents the response value for each observation; S_t is the set of observations for which the splitting result is true and S_f is the set of observations for which the splitting result is false.

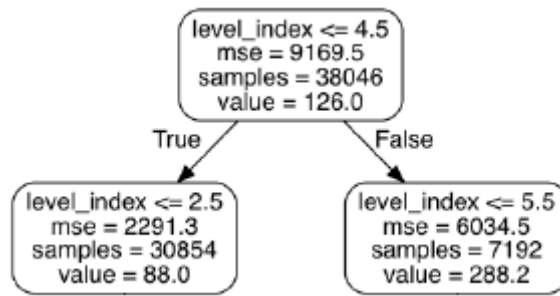


Fig.2 Sample decision tree for parametric estimation

Though decision tree method is fast (fig 2) and could be easily visualized and interpreted, the prediction accuracy is not quite favourable compared to other regression approaches. Random Forest (RF) method is thus introduced to overcome the disadvantages of decision trees [6]. Random Forest is then simply a method that grows a collection of decision trees and gives the aggregated results. What is notable in the RF model is that it also implements a smart algorithm to decorrelate the trees: Instead of searching for the best splitting feature among the whole feature space, RF model selects the best splitting feature among a random sample of m features at each internal node. Typically, m is chosen to be the approximate of square root of total number of feature space.[2] This algorithm perfectly mitigates the effect of having strong variables. Without this algorithm, decision trees in the forest will have similar structure due to the variance reduction metric and thus, the reduction in variance of the model might not be substantial.

Bayes' theorem describes the posterior or conditional probability of a hypothesis (H) based on prior knowledge of evidence (e) that might be related to the hypothesis. The posterior $p(H|e)$ of H given e is definite as:

$$p(H|e) = \frac{p(H,e)}{p(e)}$$

$$p(H, e) = p(H|e) \cdot p(e), \text{ and } p(e, H) = p(e|H) \cdot p(H)$$

$$p(H|e) = \frac{p(e|H) \cdot p(H)}{p(e)}$$

$$p(H|e) = \frac{p(e|H) \cdot p(H)}{\sum_{H^*} p(e|H^*) \cdot p(H^*)}$$

This is the Bayes' rule and lies at the core of Bayesian inference whereas H^* in the denominator is a variable that takes on all possible hypotheses

When we obtain a particular dataset D and denote θ is the parameter that we are interested in, then the posterior can be denoted as $p(\theta|D)$; the likelihood denoted as $p(D|\theta)$, which means the probability of the data might be obtained with the parameter θ under certain model assumptions; and the prior denoted as $p(\theta)$, which means the credibility of the parameter values without D [4]. The marginal likelihood or the denominator in Bayes' rule can be rewritten for continuous variables using the denotations above as:

$$p(D) = \int d\theta^* p(D|\theta^*) p(\theta^*)$$

Where θ^* denotes any possible value of θ .

Bayesian model averaging (BMA) is an application of Bayesian inferential analysis. It has been applied to model selection problems, where one combines estimation and prediction to produce a straightforward model choice criteria and less risky predictions [9]. So the average estimation across a set of models would generate more robust interval estimation, and meanwhile, reduce the type I error.

Suppose in a study, M_l is one of a set of models considered to fit the research question, Δ is the interested parameter, D is the dataset given, then the BMA-averaged Δ is the sum of specific model derived Δ /weighted by the posterior model probability $p(M_l|D)$

$$E(\Delta|D) = \sum_{l=1}^K \Delta_l p(M_l|D)$$

Although we cannot get the posterior probability $(M_l|D)$ directly, according to the Bayes' rule, the posterior for a given model M_k is:

$$p(M_k|D) = \frac{p(D|M_k) \cdot p(M_k)}{\sum_{l=1}^K p(D|M_l) \cdot p(M_l)}$$

Where $p(M_k)$ is the probability that M_k is true and the likelihood $p(D|M_k)$ is given by:

$$p(D|M_k) = \int d\theta_k p(D|\theta_k, M_k) p(\theta_k|M_k)$$

θ_k is the parameter vector of model M_k , $(\theta_k|M_k)$ is the prior density of θ_k under model M_k , and $(D|\theta_k,)$ is the likelihood.

During recent decades, Bayes methods enjoyed the popularity due to the computational progress. A class of Markov chain Monte Carlo (MCMC) algorithm became a practical method to estimate the complex random variables instead of direct sampling. A detailed tutorial is given by Hanson and Kruschke. A semi-parametric Bayesian approach and a simulation study was displayed. The computationally efficient approaches such as fully Bayesian method thus have been developed in recent years. A fully Bayesian approach for modelling and inference within GAM requires prior assumption for unknown smooth function $S(\cdot)$.

III. Proposed Model

In the proposed framework, an optimized Bayesian network estimation model is designed and implemented on the $P_{2.5}$ dataset. Initially, data normalization is applied on the air pollution data ($P_{2.5}$). Proposed data transformation method is applied on the normalized data to improve the data distribution for the joint probability estimation of the Bayesian networks. In the proposed Bayesian network classifier, a directed acyclic graph is constructed on the input transformed data. The two phase of proposed Bayesian network include, parameter estimation phase and statistical learning phase. In the parameter estimation phase, numerical or continuous attributes are estimated using the numerical parametric estimation and nominal or categorical attributes are estimated using the discrete parameter estimation measures. Finally, after parametric estimation Bayesian structure is learned using the optimized scoring function and joint probabilistic function as shown in fig 3.

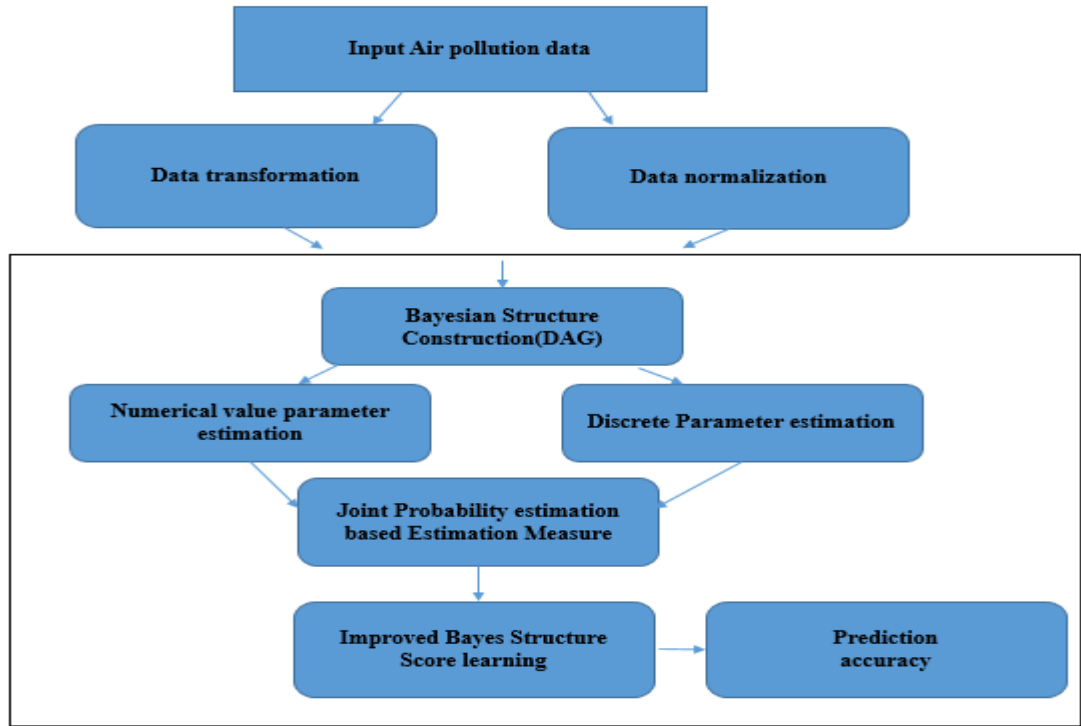


Fig. 3 Proposed Framework

III.i Optimized Bayesian Network for AQI Prediction

Optimized Bayesian network has two phases to estimate the air quality index to the given input data $P_{2.5}$. In the first phase, statistical parameters are estimated to each type of input data i.e discrete type and continuous type. In the second phase, statistical Bayesian DAG graph structure is learned using the optimized bayes scoring function. The two phases are described in the following algorithm.

Step:1: $D := \text{Input data } P_{2.5}$, A_i is the set of input random variables. I_j is the instance of attribute.

Step:2: for each attribute A in D

Step:3: Apply attribute transformation function by using following equation

$$A[i] = \left| \sum_{i=1}^n A_i^2 * (A_i - \sigma_A) / (\text{Max}(A_i) - \text{Min}(A_i)) \right|$$

Step 4: done

Step 5: Construct directed acyclic graph to the transformed data using Bayesian network.

Step 6: Computing conditional probabilities to each input variable for

joint probability estimation.

Step 7: // Phase 1

Discrete parameter estimation in the Bayesian network can be predicted using the following measure.

$$P(A_i = I_k / c_j) = \frac{N_{ijk}}{N_j}$$

Where N_{ijk} is the number of instances of class c_j having the value I_k in attribute A_i .

Step 8: Continuous parametric estimation in the Bayesian network can be estimated using the following measure.

$$P(A_i = I_k / c_j) = G(I_k, \mu_{ij}, \sigma_{ij})$$

$$G(I_k, \mu_{ij}, \sigma_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(I_k - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Here normal distribution is approximation to Gaussian distribution.

Step 9: Estimating Bayesian parameter using the traditional parameter estimation as

$$\hat{\theta}_{ijk} = \frac{N_{ijk} + \eta_{ijk}}{N_{ij} + \eta_{ij}}$$

$$\text{where } N_{ij} = \sum N_{ijk}; \eta_{ij} = \sum \eta_{ijk}$$

Step 10: //Phase 2

In this phase, statistical model selection is performed on the DAG graph to find the best fit to the input data. In the traditional Bayesian networks, constraint based and score based statistical models are used to find the best fit to the data. In the proposed approach, a novel Bayesian score is used to improve the prediction rate and to find the DAG conditional dependencies and independencies on each attribute. Basically, score based approach evaluates the dependencies and independencies in a structure that matches to the input data. This scoring measure optimizes the structure that maximizes the scoring value.

The proposed Bayesian score is computed as:

$\theta = \text{ConditionalPriorProb}(s_i);$

$\phi = \text{Joint Prob}(D / s_i);$

$\text{PropBayesScore} = \log(\theta) + \log(\phi)$

where

$$\text{Joint Prob}(D / s_i) = \prod_{i=0}^n \prod_{j=0}^{q_i} \frac{\Gamma(\sum_{k=1}^r \alpha_{ijk} \log(\alpha_{ijk}))}{\Gamma(\sum_{k=1}^r \alpha_{ijk} \log(\alpha_{ijk}) + \sum N_{ij})} \prod_{k=1}^r \frac{\Gamma(\sum_{i,j} \alpha_{ijk} \log(\alpha_{ijk}))}{\Gamma(\sum_{i,j} \alpha_{ijk} \log(\alpha_{ijk}) + \sum N_{ij})}$$

Step 11: After computing the Bayesian score to each random variable, its class and attribute probability estimations are performed based on the following steps.

for each class in classlist

do

for each node in BN attribute list

do

if(node[i]==Class)

iCP=iCP*|C|+CIndex;

else

iCP=iCP*|P_A|+P_AIValue

Compute cubic attribute distribution measure to each input attribute node and class node as

if(node[i]==Class)

Class cubic Probability =CCProb= $\sqrt[3]{\text{Node Prob}(C\text{Index})}$;

else

Attribute Cubic Probability =ACPR= $\sqrt[3]{\text{Node Prob}(P_A \text{IValue})}$;

endfor

Class Probability = CP[CIndex] = Γ CCProb

Attribute Probability = AP[P_AIValue] = Γ ACPR

endfor

Step 12: Bayesian scores and computed class and attribute probabilities are used to choose the optimal Bayesian random dependency and independency variable that fits the structure to the data.

IV. Experimental results

Experimental results are performed on the P_{2.5}delhi air polluted data taken from <https://en.tutiempo.net/climate/01-2015/ws-421820.html> and [1]. Experimental results are simulated by using java programming environment with third party libraries such as colt, jama, weka etc. Sample input data is summarized in table 3.

Input Dataset

T: Average Temperature
Tm: Minimum temperature
TM: Maximum temperature
SLP: Atmospheric pressure at sea level
H: Average relative humidity
V: Average wind speed
VV: Average visibility
VM: Maximum sustained wind speed

```
@relation P2.5
@attribute T numeric
@attribute TM numeric
@attribute Tm numeric
@attribute SLP numeric
@attribute H numeric
@attribute VV numeric
@attribute V numeric
@attribute VM numeric
@attribute 'PM 2.5' {0,1}
@data
0.398119,0.214485,0.529052,-0.467532,0.435897,-0.609195,-0.691667,-0.649932,0
-0.454545,-0.18663,-0.461774,0.564935,0.794872,-0.977011,-0.941667,-0.90502,1
0.949843,0.961003,0.865443,-0.571429,-0.717949,-0.609195,-0.2,-0.449118,1
0.586207,0.771588,0.712538,-0.525974,-0.205128,-0.517241,-0.25,-0.649932,0
-0.673981,-0.582173,-0.743119,0.837662,0.615385,-0.908046,-0.741667,-
0.845319,1
0.360502,0.481894,0.229358,-0.188312,-0.333333,-0.609195,-0.508333,-0.649932,1
0.040752,0.158774,0.003058,0.305195,0.025641,-0.724138,-0.508333,-0.649932,1
0.291536,0.392758,0.333333,0.220779,-0.102564,-0.609195,-0.125,-0.554953,0
0.122257,0.370474,0.198777,0.396104,0.051282,-0.517241,-0.275,-0.603799,0
0.015674,0.086351,0.100917,0.441558,0.307692,-0.609195,-0.433333,-0.649932,1
0.153605,0.325905,0.100917,0.435065,0.307692,-0.724138,-0.941667,-0.796472,1
0.517241,0.398329,0.559633,-0.487013,0.153846,-0.609195,-0.775,-0.796472,0
0.373041,0.303621,0.590214,-0.363636,0.564103,-0.609195,0.216667,-0.348711,0
```

0.924765,0.883008,0.865443,-0.590909,-0.461538,-0.609195,-0.258333,-0.449118,0
 0.962382,0.988858,0.840979,-0.896104,-0.564103,-0.609195,-0.541667,-0.603799,0
 0.617555,0.576602,0.437309,-0.396104,0.153846,-0.609195,-0.758333,-0.845319,0
 0.53605,0.559889,0.327217,-0.084416,-0.666667,-0.632184,-0.358333,-0.297151,1
 0.197492,0.281337,0.070336,0.220779,-0.487179,-0.448276,0.033333,-0.348711,0
 0.492163,0.409471,0.51682,-0.75974,0.435897,-0.609195,-0.6,-0.649932,0
 0.423197,0.348189,0.431193,-0.025974,0.358974,-0.609195,-0.816667,-0.845319,0
 0.398119,0.29805,0.492355,-0.474026,0.589744,-0.609195,-0.8,-0.845319,0
 0.23511,0.320334,0.094801,0.253247,-0.615385,-0.448276,0,-0.253731,0
 -0.310345,-0.175487,-0.314985,0.597403,0.384615,-0.862069,-0.45,-0.554953,1
 -0.021944,0.047354,-0.021407,0.435065,0.128205,-0.609195,0.275,-0.348711,0
 -0.680251,-0.454039,-0.547401,0.941558,0.564103,-0.908046,-0.783333,-
 0.750339,1
 0.479624,0.537604,0.69419,-0.493506,0.512821,-0.678161,-0.675,-0.649932,0
 -0.216301,-0.181058,-0.100917,0.571429,0.615385,-0.931034,-0.925,-0.90502,1
 0.887147,0.793872,0.840979,-0.694805,-0.384615,-0.425287,-0.291667,-0.449118,0
 0.504702,0.409471,0.565749,-0.194805,0.25641,-0.609195,-0.416667,-0.449118,0
 -0.166144,0.030641,-0.229358,0.694805,0.076923,-0.793103,-0.875,-0.845319,1
 0.774295,0.849582,0.700306,-0.5,-0.717949,-0.609195,-0.458333,-0.603799,1
 -0.454545,-0.181058,-0.443425,0.448052,0.615385,-0.862069,-0.816667,-
 0.750339,1
 0.38558,0.35376,0.541284,-0.25974,0.410256,-0.609195,-0.233333,0.356852,0
 0.448276,0.426184,0.406728,-0.181818,-0.307692,-0.632184,-0.108333,-0.603799,1
 -0.084639,0.08078,-0.229358,0.38961,-0.230769,-0.448276,-0.45,-0.750339,1
 -0.166144,0.08078,-0.149847,0.428571,-0.128205,-0.494253,-0.616667,-0.649932,1
 -0.304075,-0.18663,-0.406728,0.655844,0.102564,-0.563218,-0.366667,-0.554953,1
 0.090909,0.153203,0.027523,0.155844,0.025641,-0.563218,0.525,-0.253731,0
 -0.510972,-0.370474,-0.584098,0.792208,0.128205,-0.678161,-0.566667,-
 0.698779,1
 0.492163,0.543175,0.321101,-0.038961,-0.487179,-0.448276,-0.033333,-0.297151,0

IV.i Sample Normalized Data

Following data illustrates the sample normalized data to the input raw dataset. Each value in the attribute is normalized between -1 to 1 to estimate the class severity.

```
0.398119,0.214485,0.529052,-0.467532,0.435897,-0.609195,-0.691667,-0.649932,0
-0.454545,-0.18663,-0.461774,0.564935,0.794872,-0.977011,-0.941667,-0.90502,1
0.949843,0.961003,0.865443,-0.571429,-0.717949,-0.609195,-0.2,-0.449118,1
0.586207,0.771588,0.712538,-0.525974,-0.205128,-0.517241,-0.25,-0.649932,0
-0.673981,-0.582173,-0.743119,0.837662,0.615385,-0.908046,-0.741667,-
0.845319,1
0.360502,0.481894,0.229358,-0.188312,-0.333333,-0.609195,-0.508333,-0.649932,1
0.040752,0.158774,0.003058,0.305195,0.025641,-0.724138,-0.508333,-0.649932,1
0.291536,0.392758,0.333333,0.220779,-0.102564,-0.609195,-0.125,-0.554953,0
0.122257,0.370474,0.198777,0.396104,0.051282,-0.517241,-0.275,-0.603799,0
0.015674,0.086351,0.100917,0.441558,0.307692,-0.609195,-0.433333,-0.649932,1
```

IV.ii Sample Transformed data

The following data describes the transformed values on the normalized data. This process is required to improve the data distribution in order to estimate joint probability estimation.

```
8.810391,4.445678,8.687254,88.751602,0.109004,150.22566,94.162254,27.419424,
0
1.904632,3.152431,6.467221,79.541337,61.701872,154.730389,73.549949,169.6858
17,1
60.603477,33.084415,78.106908,11.999393,55.102636,118.693337,129.890304,199.
752722,1
69.580987,33.084415,69.151919,7.394261,46.303655,127.7026,157.373433,179.151
332,1
84.773636,61.138795,98.083343,19.469071,26.505948,141.216396,116.148823,159.
663653,1
41.267439,25.48641,49.175483,39.630188,33.105184,141.216396,30.885674,97.859
277,0
```

107.562664,80.426142,118.059779,35.587094,24.306202,163.739652,149.128511,1
89.173702,1
2.929378,15.426131,16.111029,62.655851,77.100089,118.693337,61.182566,106.76
8142,0

IV.iii Sample Bayesian estimated test result

The following result illustrates the proposed Bayesian network estimation to each sample in the test instances. Each test instance is evaluated using the joint probability estimators and Bayesian score to predict the class using the DAG dependencies and independencies.

23.646656,42.896049,8.687254,111.777265,96.897883,150.22566,103.78144,137.94
8552,1 =====> predicted class index value :0.0
79.939626,33.668877,71.907308,23.512165,11.107731,177.253449,190.35312,219.2
40607,1 =====> predicted class index value :1.0
90.988885,62.307719,102.905191,9.491343,17.706967,186.262712,157.373433,189.
173702,1 =====> predicted class index value :1.0
21.574906,19.517469,24.530662,86.449095,50.703146,150.22566,98.284715,137.94
8552,0 =====> predicted class index value :0.0
31.933545,30.622244,29.352622,81.076381,94.698137,118.693337,127.141942,159.
663653,0 =====> predicted class index value :0.0
77.177366,41.266881,78.106908,27.349835,17.706967,177.253449,160.12163,139.3
35239,1 =====> predicted class index value :1.0
82.701886,54.709609,106.3494,2.583585,15.507221,163.739652,145.00605,179.151
332,1 =====> predicted class index value :1.0
131.732822,86.855222,131.147797,36.354557,8.907985,172.748915,151.876708,16
9.685817,1 =====> predicted class index value :1.

Table.3 Performance analysis of traditional and proposed probabilistic estimation models on input data

Model	MAE	RMSE	R-square
ANN	0.56	0.81	0.42
ARIMA	0.52	0.73	0.59
SVM	0.63	0.65	0.48
Bayesian network	0.46	0.52	0.78
Proposed Model	0.26	0.15	0.84

Table 3, describes the performance analysis of present approach to the traditional approaches for AQI prediction. From the table, it is noted that the present approach has less error rate and high accuracy compared to the traditional approaches.

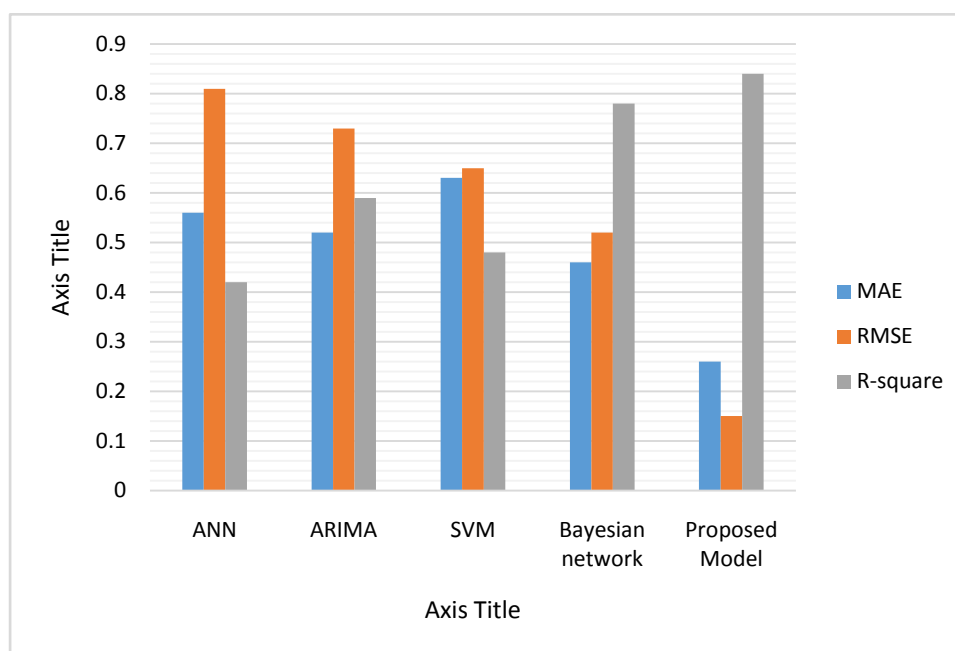


Fig. 4 Performance analysis of traditional and proposed probabilistic estimation models on input data

Figure 4, describes the performance analysis of present approach to the traditional approaches for AQI prediction. From the figure, it is noted that the present approach has less error rate and high accuracy compared to the traditional approaches.

V. Conclusion

Air quality information is gathered remotely using air quality monitoring sensors which are outfitted with a variety of vaporous. IOT distribution mechanisms $PM_{2.5}$ is one of the major factor to predict the air quality index(AQI) and its severity level. Due to high noisy and outliers in the $PM_{2.5}$ data, it is difficult to classify and predict the air quality by using the traditional quality prediction models. In order to overcome these issues, an optimized Bayesian networks based probabilistic inference model is designed and implemented on the air quality data. This paper proposed improved discrete and continuous parameter estimation and bayes score optimization are implemented on the air quality prediction process. Experimental results show that the present optimized Bayesian network classifies and predicts the air quality data with high less computational time and accuracy.

References

- I. Ayaskanta Mishra, "Air Pollution Monitoring System based on IoT: Forecasting and Predictive Modeling using Machine Learning", International Conference on Applied Electromagnetics, Signal Processing and Communication (AESPC), 22nd - 24th October-2018, Bhubaneswar, Odisha, India, IEEE, Paper ID# 9.
- II. C. Li and Z. Zhu, "Research and application of a novel hybrid air quality early-warning system: A case study in China", *Science of The Total Environment*, vol. 626, pp. 1421-1438, 2018. Available: 10.1016/j.scitotenv.2018.01.195 [Accessed 20 February 2019].
- III. Hybrid improved differential evolution and wavelet neural network with load forecasting problem of air conditioning *Int. J. Electr. Power Energy Syst.* 61, 673–682
- IV. H. Li, J. Wang, R. Li and H. Lu, "Novel analysis–forecast system based on multi-objective optimization for air quality index", *Journal of Cleaner Production*, vol. 208, pp. 1365-1383, 2019. Available: 10.1016/j.jclepro.2018.10.129 [Accessed 20 February 2019].
- V. https://raw.githubusercontent.com/alyakhtar/AQI-Delhi/master/Data/Original-Data/Original_Combine.csv
- VI. K. Gan, S. Sun, S. Wang and Y. Wei, "A secondary-decomposition-ensemble learning paradigm for forecasting $PM_{2.5}$ concentration", *Atmospheric Pollution Research*, vol. 9, no. 6, pp. 989-999, 2018. Available: 10.1016/j.apr.2018.03.008 [Accessed 20 February 2019].

- VII. S. Feng, F. Jiang, Z. Jiang, H. Wang, Z. Cai and L. Zhang, "Impact of 3DVAR assimilation of surface PM 2.5 observations on PM 2.5 forecasts over China during wintertime", *Atmospheric Environment*, vol. 187, pp. 34-49, 2018. Available: 10.1016/j.atmosenv.2018.05.049 [Accessed 20 February 2019].
- VIII. T. Fontes, P. Li, N. Barros and P. Zhao, "A proposed methodology for impact assessment of air quality traffic-related measures: The case of PM2.5 in Beijing", *Environmental Pollution*, vol. 239, pp. 818-828, 2018. Available: 10.1016/j.envpol.2018.04.061 [Accessed 20 February 2019].
- IX. Wang, J., Hu, J., 2015. A robust combination approach for short-term wind speed forecasting and analysis - Combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (Extreme Learning Machine), SVM (Support Vector Machine) and LSSVM (Least Square SVM) forecasts using a GPR (Gaussian Process Regression) model. *Energy* 93, 41–56.
- X. World Health Organization, "Monitoring ambient air quality for health impact assessment," WHO Regional Office Eur., Copenhagen, Denmark, Tech. Rep. 85, 1999.
- XI. World Health Organization. Occupational and Environmental Health Team. (2006). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment. Geneva: World Health Organization. <http://www.who.int/iris/handle/10665/69477>.
- XII. Yuan, X., Tan, Q., Lei, X., Yuan, Y., Wu, X., 2017. Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine.
- XIII. Y. Cheng, H. Zhang, Z. Liu, L. Chen and P. Wang, "Hybrid algorithm for short-term forecasting of PM2.5 in China", *Atmospheric Environment*, vol. 200, pp. 264-279, 2019. Available: 10.1016/j.atmosenv.2018.12.025 [Accessed 20 February 2019].