An efficient feature selection technique using genetic algorithm for air pollution prediction

Gagandeep Kaur
M.Tech, Computer Science & Engineering,
Sri Sai college of Engineering and Technology
Manawala, Amritsar

Harmenpreet Kaur
Asst. Professor, Computer Science & Engineering,
Sri Sai college of Engineering and Technology
Manawala, Amritsar

***Abstract*—The using benzene(C6H6) in mining contains the possible ways to affect the products surrounding surface benzene(C6H6) as well as groundbenzene(C6H6). In answer to ecological concerns as well as government policy, the mining industry worldwide more and more monitors benzene(C6H6) discharged from mine sites, as well as possesses implemented a variety of organization strategies to avoid air pollution.As a result, it is necessary to monitor the quantity and quality of benzene(C6H6). Data mining techniques are used to predict the air pollution. This paper has proposed the integration of the feature selection technique and Genetic algorithms to improve the accuracy rate for detection of air pollution as well as substantial features of air pollution. The experimental results brings about the proposed technique that clearly shown the fact that proposed technique outperforms over the existing methods.**

Keywords*— *Data mining, Benzene(C6H6) mining, Air pollution,Feature Selection and Genetic Algorithm

I. INTRODUCTION

Use of data mining is increasing in air pollution area in these days. As we know various applications of air pollution data mining are gaining popularity day by day. Mainly data mining concept is basically used in air pollution management and classification analysis to improve the success rate. Main problem faced by the air pollution sector is to improve the performance of data mining algorithms by using valuable technique. Air pollution falls under respective classes. The quality of benzene(C6H6) may be good, fairly good, poor or the quality of benzene(C6H6) can be grossly polluted. Air pollution is now becoming a significant matter due to the raising citizenry size. As the populace raises which generates higher possibility for harmful substances that can invade into benzene(C6H6) supplies. When a substance enter into a body of benzene(C6H6) that leads to various problems. Benzene(C6H6) pollution is considered to be harmful and thus there is a need to reduce the benzene(C6H6) pollution so that the air pollution can be easily maintained. As a result, it is necessary to monitor the quantity and quality of benzene(C6H6). Air pollutiondegradation caused by erosion and sedimentation and agricultural runoff threaten its fish and wildlife. Hence, air pollution has become a greater priority nowadays. Domestic waste, sewage, wastebenzene(C6H6) and overuse of benzene(C6H6) continue to lower the benzene(C6H6) level. Air pollution is investigated at different locations and samples are collected to test them in a laboratory for analysis whether benzene(C6H6) is clean benzene(C6H6) and is intended to use for drinking purposes.Climate, temperature and vegetation are some of the important natural conditions that adversely influence the air pollution.Most of the researchers studied the classification models and hence made an comparison between ANN and Bayesian network to predict the air pollution [13]. Lots of the reports have already been done on air pollution and benzene(C6H6) pollution. In order to protect the health of living organisms, various controlling activities must be taken into account that pollute them.

1.1 Machine Learning in Support Vector Machines

Machine learning is the subfield of computer science that offers computers the capability to learn without having to beexplicitly programmed.Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is unfeasible. For example spam filtering, optical character recognition (OCR) search engines and computer vision.In support vector machines, data point is viewed as p-dimensional vector (a list of p numbers).Consider there are some data points and they each belong to one of the two classes and we want to know that to which class the new data point will belong to and we also want to know that these data points can be separated by (p-1) hyperplanes.This is called as linear classifier. There are many hyperplanes that might classify the data. The best hyperplane is the one which provides the largest separation between the two classes.The hyperplane that is having the maximum distance from it to nearest data point on each side is chosen. The hyperplane is known as the maximum-margin hyperplane and the linear classifier is known as a maximum margin classifier.

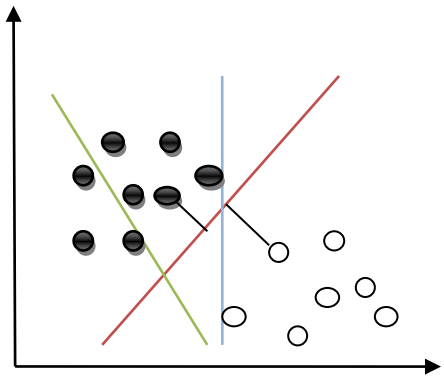

Fig. 1. Depiction of Hyperplanes

H_1-It does not separate the classes.
H_2-It separates the classes but with small margin.
H_3-It separates the classes with maximum margin.

1.2Air pollution

Benzene(C6H6) is the lifeblood of the environmental surroundings which is important for the survival of all residing things- plant, animal and individual and we ought to do everything probable to steadfastly keep up its quality for nowadays and the future. Several facets influenceair pollution. Elements contained in the air will significantly influence rainfall. Dirt, volcanic gases, and organic gases in the air, such as for example carbon dioxide, oxygen, and nitrogen, are all typical contained in rain. When differentelements such as for example sulphur dioxide, hazardous chemicals, or lead have been present in the air, they are also gathered in the torrential rain since it comes to the ground. Yet another component influencing air pollution could be the runoff from metropolitan areas. It will acquire the trash littering the roads and bring it to the obtaining supply or benzene(C6H6) body. Metropolitan runoff exacerbate the air pollution in streams and ponds and seas by raising the levels of such elements as nutritional elements (phosphorus and nitrogen), residue, animal wastes (fecal coliform and pathogens), oil products, and street salts. Each one of these outcomes might have an adverse effect on the marine environment and producebenzene(C6H6) unsuitable for recognized or possible uses.

Air pollution can be regarded as the physical, chemical and biological properties that address its suitability for a designated use. It determines the appropriate concentration of chemicals present in the benzene(C6H6). Various technologies have been used to remove the harmful contaminants from the benzene(C6H6) before it is used for different purposes. A PSO based artificial neural network model and many other methods have been studied and analyzed that shows effective evaluation and prediction applications. Many researchers discussed various techniques that can classify and predict the air pollution. This results in the reduction of benzene(C6H6) pollution which leads to benzene(C6H6)-borne diseases.

II. PROPOSED TECHNIQUES

1. Feature Selection:

Feature selection, named as variable selection, attribute selection or variable subset selection, is the method of choosing a part of relevant features (variables, predictors) for utilize in design construction. Feature selection attempts to choose the simply sized subset of features depending on the following. It can be:

1. The classification accuracy does not considerablyreduce as well as
2. The resulting class allocation, knownonly the values for the picked features, can be as close as possible to the unique class allocation, known all features.

There are lots of potential great things about variable as well as feature selection: making ease of data visualization as well as data understanding, lowering the measurement and storage requirements, reducing training as well as utilization times, defying the problem of dimensionality to boost prediction performance.

2. Genetic Algorithm:

Genetic Algorithms are certainly one of computational models stimulated together with the development and natural Selection processes. They design the most perfect option is together with the problem right data structure called chromosome**,** genotype or genomesignifying the wide ranging solutions, called individuals or creatures or phenotypes. Quite a few genetic operators are provided about bat roosting chromosomes to acquire a quite high optimization together with the problem. An innate algorithm begins employing a arbitrary population of individuals. Up to date inhabitants are changed with the genetic operators' right new generation of individuals. The key element goal of a period may be to keep your best individuals, improving the fitness together with the population, until some stopping criteria is achieved. This criterion is usually a fitness threshold, the sheer quantities of generations or lacking improvement.

1. Selection operator's makes using the evaluation function to find out what humans possess the biggest potential. They might persist whilst in the population and used with the other operators.
2. The recombination operators (mutation and crossover) are traditionally utilized to produce new individuals using a few high potential individuals. They are designed to diversify the search process. The widely accepted operators of these kinds are cross-over as well as mutation.
3. The cross-over operator uses 2 if not more fractions of high potential individuals to develop a fresh individual that's appended to your higher generation with all the population.
4. The mutation operator, on contrary, takes one high potential individual as well as really a little alternation in amongst its components. The modern person is usually appended over the next generation with all the population.

III. RELATED WORK

S. Sasikala et.al [1] developed a memetic algorithm by adding Genetic algorithm and Shapely value for multi-class classification named as SVEGA through managing various dimensional data. A SVEGA model is applied as the feature selection instrument which serves for the improvement of classification engine. Mark E. Borsuk et.al [2] proposed a BOD decay approach which employs Bayes' theorem to generate a joint probability distribution for several parameter values contingent on the seen data. Applying this jointdistribution, a mathematical integration method is utilized to obtain limited parameter distributions which can be applied to directly evaluating the general plausibility of competitive parameter values. Zexuan Zhu et.al [3] discussed a novel hybrid filter and wrapper feature selection algorithm that employs filter ranking approach as LS heuristic. The exploratory outcomes shown reveal that the planned technique queries more proficiently and is effectiveat providing excellent classification accuracy with a couple of characteristics simultaneously. Salisu Yusaf Muhammad et.al [4]developed an appropriate classification model for analyzing as well as classifying air pollution in accordance with the machine learning algorithms.Shah Christirani Azhar et.al [5] has discussed three multivariate mathematical practices (CA, PCA, and DA) which aids to identify nine tracking programs situated on the stream into categories of relatedair pollution traits predicated on six pickedair pollution variables.This purpose possessed a appropriate classification performance of 100%. JanetR. Barclay et.al [6] recommended that the associations between air pollution and land cover may potentially increase benzene(C6H6)body classification techniques, and we suggest improved defenses for these Class Type B benzene(C6H6)sheds to keep up their AA-like air pollution for possible future drinking benzene(C6H6) supply. ZhenXiang Xing et.al [7] proposed a fuzzy comprehensive evaluation model basically depends on entropy weight method (FCE-EW) that was created to measure the precise state condition of underground air pollution. K.Z. Mao et.al [8] proposed a SVM discriminative function pruning analysis (DFPA) algorithm for featuresubset selection. The potency of the DFPA strategies have been approved applying two big real life problems. Yaonan Wang et.al [9] developed a middle initialization technique centered on MST to deal with the issue of regional minima. Contrast results display that our initialization approach is preferable to or equivalent with CCIA andkd-tree with regards to sample acceptance rate but computationally costly in certain cases. S. Wechmongkhonkon et.al [10] discussed a MLP neural network utilizing the Levenberg-Marquardt algorithm is employed to categorize the air pollution of Dusit district canals of Bangkok, Thailand. The outcomes show that the neural system accomplish a higher accuracy classification percentage of 96.52%. Adam Woznica et.al [11] proposed a general framework that focused on various feature designs made from a certain dataset that are able to remove designs from within the models. He found that the closedApr aggregation method that gives an excellent compromise involving the stability andpredictive performances. The allMed strategy that depends on the k-medoids clustering algorithm and hence it is an easy task to implement. Lee Yoot Khuan et.al [12] has shown the back propagation neural network, the modular neural network and the radial basis function network which have been applied to model and thus estimate the Air pollution Index. The MNN model was discovered to be the most appropriate model for use to ascertain the WQI, in relation to accuracy and rapidly understanding time. Chamara P Liyanage et.al [13] discussed various classification models for evaluating the air pollution to reveal the best accuracy. The BN model has shown the most appropriate accurate accuracy as compared to the MLP model.

IV. METHODOLOGY

5.1 Proposed Algorithm

The proposed work includes feature selection technique based genetic algorithm with various data mining algorithms.

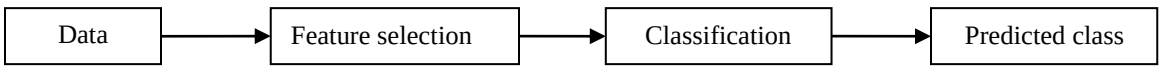

Fig 2: Research Framework

The metrics vectors contain the form: $x_i G = [x_{1,i,G}, x_{2,i,G,\ldots} x_{D,i,G}] = 1,2 \ldots, N$ where G is the generation number.

Step1. Initialization:

1. Initialize upper as well as lower bounds for every metric:

$$x_J^L \leq x_{j,i,1} \leq x_J^L$$

2. Arbitrarilychoose the initial metric values consistently on the intervals $[x_J^L, x_J^L$

Step2. Mutation: Mutation is a process in which a bit involves flipping it, changing 0 to 1 and vice-versa. For example:

Before	0	1	1	1
After	0	1	0	1

1. Every N metric vectors go through mutation, crossover as well as selection.
2. Mutation broad the search space.
3. For a known metric vector x_{iG} arbitrarily choose three vectors$x_{i1,G}$, $xr2,G$, as well as $xr3,G$, for example which indices $i,r1,r2$ and $r3$ are different.
4. Include the weighted variation of two of the vectors to the third $v_i,G+1=x_r1,+1G+(x_r2,G-x_r2,G)$
5. The mutation factor F is a constant from [0, 2].
6. v_iG+1 Is called the donor vector.

Step 3: Crossover: Crossover is a process of taking more than one parent solutions and producing a child solution from them. For example:

Parent 1
Parent 2

A	B	C	D	E	F	G	H
H	G	F	E	D	C	B	A

H	B	C	D	E	G	F	A

Offspring

1. Crossover incorporates achievable solutions from the previous generation.
2. The trial vector u_iG+1 is constructed from the elements of the target vector,x_{iG} as well as the elements of the donor vector, v_iG+1
3. Elements of the donor vector enter the trial vector with probability CR.

$$u_{ji}G+1=\begin{cases}u_{ji}G+1\ if Rand_{ji}\le CRorj1_{rand}\\ x_{ji}G+1 if Rand_{ji}\ne CRorj1_{rand}\end{cases}$$

$$i=1,2\dots\dots N;j\ 1,2,\dots\dots D$$

$Rand_{ji}\sim U[0,1]1_{rand}$is a random integer from $[1,2,\dots,\mathrm{D}]$

$$1_{rand}\text{ensures that}u_{ji}G+1\ne x_iG$$

Step 4: Selection: Selection is a process that gives preference to improve individuals, permitting them to transfer their genes or individuals to a higher generation. The goodness of every individual is dependent upon its fitness. Fitness may be based upon goal function or using a subjective judgment.

1. The target vector x_iG is compared with the trial vector v_iG+1 and the one with the lowest function value is admitted to the next generation

$$u_{ji}G+1=\begin{cases}u_{ji}G+1\ iff(u_{ji}G+1)\ \le x_iG\\ \qquad i=1,2,\dots\dots N\\ x_iG+1\text{otherwise}\end{cases}$$

2. Mutation, crossover as well as selection continue unlessa number of stopping criterion is reached.

In proposed algorithm first it will check what the problem for which it will work is. Next we will collect the data that is required. We have collected air pollution dataset from the respective site air pollutionarff for air pollution classification and for prediction of the air pollution class. In next step data selection and transformation is performed. Feature selection technique is helpful for selecting the desired attributes from the list of attributes. Once dataset is loaded we apply the data mining algorithms such as Naïve Bayes, Decision Table, Random Forest etc for air pollution classification.

5.2 Proposed Methodology:

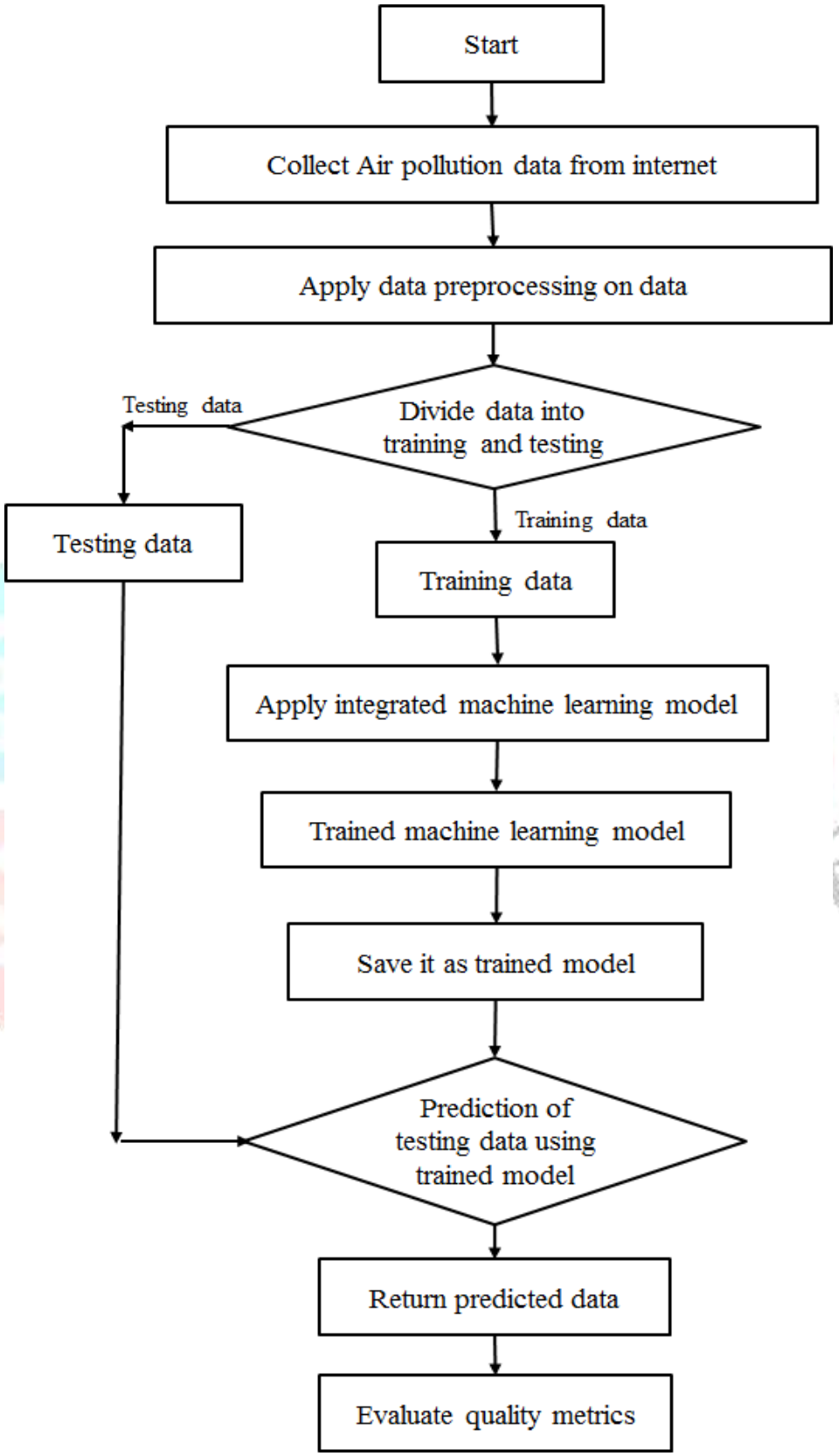

Fig 3: Flow Chart of Proposed Methodology

V. ANALYSIS OF RESULTS

This section covers the various data mining algorithms using genetic algorithm between existing and proposed techniques. The projected algorithm is tested on several data mining algorithms. Some familiar accuracy parameters for air pollution have been chosen to show that the efficiency of the projected algorithm is better than the existing techniques. For experimentation and implementation the proposed technique named as genetic algorithm is evaluated using MATLAB tool u2013a. Here we will compare the performance of existing data mining algorithms with the feature selection technique based on genetic algorithm evaluate the parameters TP rate, FP rate, Precision, Recall and F-measure.

The results are tabulated in the respective table given below. The air pollution dataset the features are reduced from the large dataset. The feature selection technique will select only those attributes which are relevant for air pollution classification. We have total 30 attributes in the dataset and only 9 attributes are selected for air pollution data mining process. This reduced dataset is then used for air pollution classification where the "numerictonominal" attribute is applied before the classification. The dataset is loaded to classify the air pollution and the results are recorded by evaluating the various parameters.

Table 1.Feature selection based Genetic algorithm for air pollution

Dataset	No. of attributes	No. of classes
Air pollution (without GA)	30	4
Air pollution (with GA)	9	4

1. TP Rate:

TPR refers to True Positive Rate and defined as measurement of positive cases which are properlyrecognized. It is prediction of correctly identified instances.TPR can be expressed by using formula:

$$\text{TP Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2. FP Rate:

FPR is called False Positive Rate and defined as portion of those instances or objects which are imperfectlyrecognized as positive. FPR can be expressed by using the formula:

$$\text{FP Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

3. Precision:

Precision is defined as measurement of all positive cases that are identified when making calculations. Higher Precision signifies that an algorithm significantly returned more significant results when compared to irrelevant. Precision can be calculated by using the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

4. Recall

Recall is the division of the written documents which are appropriate to the query that have been convincingly recovered. It is also called TP Rate or Sensitivity. It is defined as collection of positive cases. Recall can be expressed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

5. F-Measure

It contains together precision as well as recall. It is usually utilized to check the accuracy as well as reliability. It computes the mean of precision and recall. Basically, it uses as best and 0 as worst when both precision and recall are used.

F-measure can be calculated with using the formula given as:

$$\text{F} - \text{Measure} = 2 * \frac{\text{P} * \text{R}}{\text{P} + \text{R}}$$

6. ROC Area

It is a just a frequently used graph that summarizes the efficiency of a classifier over-all probable thresholds. It is mainly produced by representing the TP Rate(y-axis) contrary to the FP Rate (x-axis)as it vary the threshold for assigning findings to a certain class. Basically, it uses as best and 0 as worst when both precision and recall are used. It is commonly referred to as Receiver Operating Characteristics.

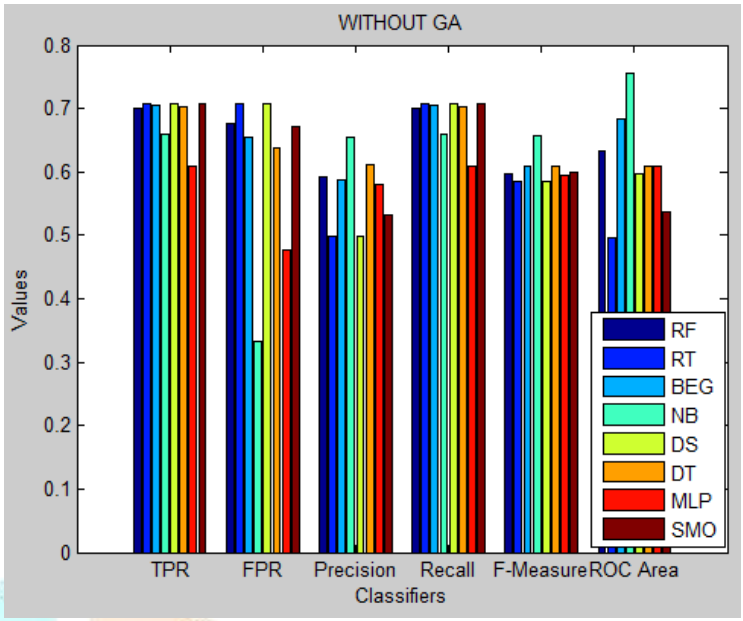

Fig 4: Comparison of TPR, FPR, Precision, Recall, F-Measure and ROC area by each classifier without GA

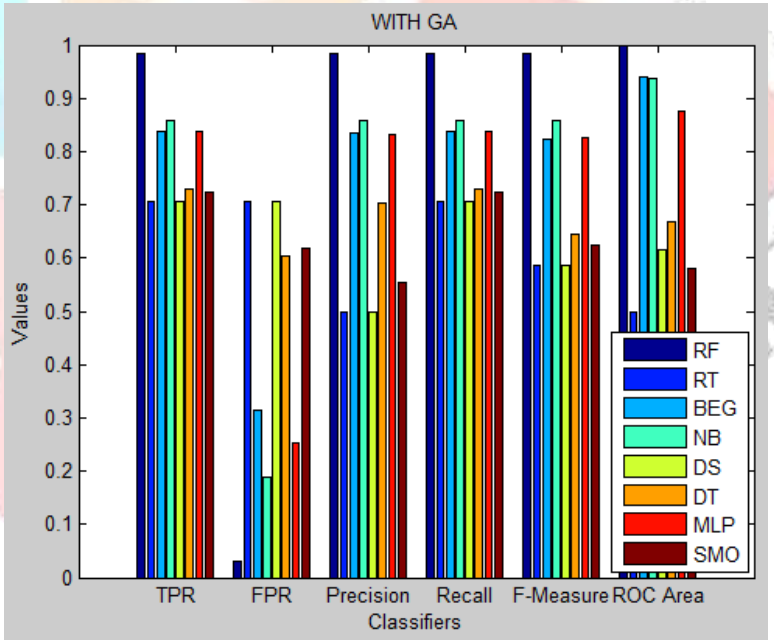

Fig 5: Comparison of TPR, FPR, Precision, Recall, F-Measure and ROC area by each classifier with GA

Fig 5 represents that the random forest with genetic algorithm outperforms because it shows maximum output for classifying the air pollution.

VI. CONCLUSION

The usage of benzene(C6H6) in mining has got the potential to influence the quality of surrounding surface benzene(C6H6) and groundbenzene(C6H6). In this paper it represents that existing literature has introduced the detection of air pollution has been considered and evaluates by using various state-of-the art algorithms but still there are some issues left to enhance the accuracy rate further for recognition of air pollution.So to improve this new method has been proposed i.e. by integrating the feature selection technique and Genetic algorithmsthat enhances the accuracy rate of air pollution as well assubstantial features of air pollution.The proposed technique has been designed and implemented in weka and Matlab toolbox 2013. The simulation result shows that by

applying integration of the feature selection technique and Genetic algorithms by using various parameters i.e. TP Rate, FP Rate, Precision, Recall, F-Measure and ROC Area. This proposed method shows better results as compared to existing results.

REFERENCES

[1] S. Sasikala, S.A.A. Balamurugana, S.Geetha, "A novel adaptive feature selector for supervised classification", Information Processing Letters 117 (2017) 25–34.
[2]M.E. Borsuk and C.A. Stow, "Bayesian parameter estimation in a mixed-order model of BOD decay," Benzene(C6H6) Research,vol.34(6),pp.1830-1836, 2000.
[3] Z. Zhu, Y.S. Ong, M. Dash, Wrapper–filter feature selection algorithm using a memetic framework, IEEE Trans. Syst. Man Cybern., Part B 10(4) (2006) 392–404.
[4] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. A. Aziz, A. A. Jamal, "Classification Model for Air pollution using Machine Learning Techniques", International Journal of Software Engineering and Its Applications Vol. 9, No. 6 (2015), pp. 45-52.
[5] S.C. Azhar, A.Z. Arisa, M.K.Yusoffa, M.F. Ramli, H. Juahir, "Classification of river air pollution using multivariate analysis", International Conference on Environmental Forensics 2015 (iENFORCE2015).
[6] J.R. Barclay, H. Tripp, C.J. Bellucci, G. Warner,A.M. Helton, "Do benzene(C6H6)body classifications predict air pollution?", Journal of Environmental Management 183 (2016) 1-12.
[7] Z. Xing, Q. Fu, D. Liu, "Air pollution Evaluation by the Fuzzy Comprehensive Evaluation based on EW Method", 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).
[8] K.Z. Mao, "Feature subset selection for support vector machines throughdiscriminative function pruning analysis,"IEEE Trans. Syst., Man,Cybern. B, Cybern., vol. 34, no. 1, pp. 60–67, Feb. 2004.
[9] Y. Wang, C. Li and Y. Zuo, "A Selection Model for Optimal Fuzzy Clustering Algorithm and Number of Clusters Based onCompetitive Comprehensive Fuzzy Evaluation", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 17, NO. 3, JUNE 2009.
[10] S. Wechmongkhonkon, N.Poomtong, S. Areerachakul, "Application of Artificial Neural Network to Classification Surface Air pollution," International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering Vol:6, No:9, 2012.
[11] A. Woznica, P. Nguyen, A. Kalousis, "Model mining for robust feature selection", KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM New York, NY, USA, PP 913-921, 2012.
[12] L. Khuan, N. Hamzah and R. Jailani, "Prediction of Air pollution Index(WQI) Based on Artificial Neural Network(ANN)",Conference onResearch and Development Proceedings, Malasia, 2002, pp. 157-161.
[13] C.P. Liyanage, K. Yamada, "Comparison of Air pollution Classification Models Using Artificial Neural Network and Bayesian Network", 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems.
[14] S.W. Athukorala, L.S. Weerasinghe, M. Jayasooria, D. Rajapakshe,L.Fernando, M. Raffeeze, N.P. Miguntanna, "Ananlysis of Air pollution Variation In Kelani River, Sri Lanka Using Principal ComponetAnalysis," SAITM Research Symposium on Engineering Advancements2013.
[15] T. Wijesinghe, "Status of Air pollution of Kelani River, CentralEnvironmental Authority, Sri Lanka," IIIRR online publication pp.255,2010.
[15]Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, "Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshbenzene(C6H6) and estuarine reservoirs, Korea.," Sci. Total Environ., vol. 502, pp. 31–41, Jan. 2015.
[16] S. Maiti and R. K. Tiwari, "A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundbenzene(C6H6) level prediction," Environ. Earth Sci., vol. 71, no.7, pp. 3147–3160, 2013.
[17] M.J. Diamantopoulou, V.Z. Antonopoulos and D.M. Papamichail "The Use of a Neural Network Technique for the Prediction of Air pollution Parameters of Axios River in Northern Greece", Journal 0fOperational Research, Springer-Verlag, Jan 2005, pp. 115-125.