International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.165-174
http://dx.doi.org/10.14257/ijdta.2015.8.6.15

Integrating Normalization with Random Tree Data Mining Approach for Mining Cloud Based Big Data's of Asthma Patient's

Abhinav Hans[1] and Sheetal Kalra[2]

[1]*Department of CSE, GNDU Regional Campus, Jalandhar*
abhinavhans@gmail.Com
[2]*Department of CSE, GNDU Regional Campus, Jalandhar*
sheetal.kalra@gmail.com

Abstract

The potential of cloud computing for overriding the needs for deploying various infrastructures for running a server based services brought up a revolutionary change in the way the traditional demands of the people use to be handled. Cloud computing provides the rental service for the user in which a user can use the particular software by paying for that on the cloud server. Since the whole scenario is beneficial to big industries like Facebook, Google, Orkut etc., various other fields are also getting dependent on cloud computing. Since tons of data is uploading every second to the cloud server does need to be mined properly for efficient data storage. In this paper we try to integrate the data preprocessing technique with data classification technique to mine big data's of asthma based patients. We have used simulation tool called eclipse to run the API's of weka and cloudsim for setting up the experimental environment.

Keywords: *Big Data, Data Mining, Data pre- processing, Data Classification, Random Tree*

1. Introduction

The technology demanding less resources and with supreme output always attracts the user. So the Cloud Computing is the one which provides the same and entices the users and managing one's own servers [1]. Though there are many famous technologies like wireless sensor networks, adhoc network, but cloud computing is the most famous technology amongst all. The property by providing the software's on lease allows the cloud to overcome the other technologies. The benefit of prepaid service of cloud computing allowed the resources to be accessible by any type of the user anywhere and anytime. Various factors like scalability, low recovery cost, less maintenance, huge data storage provisions, speedy deployment and many more factors make cloud the most powerful approach. Since the cloud is associated not only with the information technology, but also with many other fields that in the human health, sales and management files too. By monitoring the present conditions of the patient by using particular body sensor network the hospital expenses can be avoided [2].

Extracting a quite useful knowledge from various numbers of data sources that gives necessary information by separating patterns, symbols, attributes *etc.* is known as data mining. Data mining is a multi-disciplinary field of large database in which any required data can be fetched and used by user's requirement. A number of phases like Data understanding, Data preparation, Modelling, Evaluation, and Deployment are the unique part of data mining. Various social media websites are totally dependent on the data mining process as the amount of data that gets uploaded every single hour is in tons of terabytes, so it makes data mining process to be very much significant. Although not any social media websites, but many different sections like hospitals, IT industries, online

ISSN: 2005-4270 IJDTA

shopping *etc.* are also bringing data mining as a significant approach in their data maintenance job.

1.1. Models of Cloud Computing

The cloud system consists of three service models based on the basis of resource requirement, *i.e*. SaaS, PaaS and IaaS [14]. Various cloud computing models that provide the facility to the user as required are discussed below.

- *SAAS (Software as a service):* Software as a service provides the software on lease *i.e*. pay per use service of software on a cloud server.
- *IAAS (Infrastructure as a service):* it provides the particular infrastructure to cloud services by means of virtual machines and other hardware requirements.
- *PAAS (Platform as a service):* To run the application on the cloud there must be an surroundings where this service must run. Therefore the marketers cater the platform where the operating system, web server, programming language execution environment is provided.
- *SEECAAS(Security as a service):*Many confidential data that user tries to hide from various internet threats, security as a service provides a huge help by providing a protocol based security on the cloud.

1.2. Asthma and its Types

Asthma diseases diagnose is totally dependent on the sound that is created by the patient during cough. A non-asthmatic patient when coughs produces a sound of frequency (206(14) Hz) where as an asthmatic patient on cough produces a sound with frequency (239(19)Hz).There are various sound recording devices now a days that records the sound and shows the frequency of the sound from which it can be clearly diagnose whether the sound is of a normal person or of an asthmatic patient. On the basis of bronchial hypersensitivity the asthma has the following types: allergic asthma (atopic, extrinsic, caused by immunologic stimulus of an antigen), intrinsic (non-allergic, induced by infection, physically or chemically), exercise induced, drug induced asthma, occupative asthma and asthmatic bronchitis [3]. There are different types of sounds according to different theories. According to the earlier American Thoracic Society, sounds are considered “continuous” if their duration is longer than 250 ms; otherwise they are considered “discontinuous” [10]. High-pitched continuous sounds (dominant frequency above 400 Hz) and rhonchi as low-pitched continuous sounds (dominant frequency of 200 Hz or less) is considered as wheeze according to the ATS.
But according to the new definition of CORSA (Computerized Respiratory Sound Analysis) guide- lines, the dominant frequency of wheeze is usually above 100 Hz and the duration greater than 100ms [10]. Wheezes are continuous adventitious sounds, which are superimposed on normal breath sounds and often associated with bronchial airway obstruction. There are many circumstances leading to wheezing which include all mechanisms narrowing airway calibre such as bronchospasm, mucosal edema, and external compression by a tumor mass, or dynamic airway obstruction [14]. Asthmas adverse effect is according to the symptoms a patient suffers. Although asthma can be classified into four stages on the basis of symptoms:

1.2.1. Intermittent: Patient suffers light cough and wheezing for less than twice per week and at night less than twice per month.

1.2.2. Mild Persistent: Patient gets an asthma attack at least once in a week. Shortening of breath, heavy cough, wheezing, chest tightness occurs.

1.2.3. Moderate Persistence: The big air passageway of the lungs is affected by this, heavy coughing in time slots and wheezing.

1.2.4. Severe Persistent: Continues episodes occur all day and night time for several days, persistent cough and wheeze.

2. Literature Survey

Asthma is the most ascent disease now days in which age factor doesn't matters *i.e*. it can be in People of any age group. In [11] author has presented stepwise the background of asthma in medical terms followed by information about the Pathology and symptoms later. After that author has highlighted some the drawbacks of the existing techniques for managing asthma by emphasizing on showing the important disease management techniques in the traditional way. A tele-monitoring technique on glide paths to asthma is done.

Asthma is a severe disease and can be very harmful effects if it's not taken seriously. By taking its serious impact on health, a continuous monitoring is must to check the body and respiration behavior of the patient. The most countable factor is the environment in which they breathe. So In [12] author proposes a development of a rule-based asthma system. So according to it, the patients are given various suggestions on the possibilities of occurring an asthma attack, according to patient's current body conditions and the environment in which they breathe. The system is based on the questioning process to the patient and answer given by patients defines the patient's present health condition and the environmental condition in which they are living with.

This research work puts light on the data mining procedure in which diabetes disease can be predicted on the basis of the medical record history of the patient. Diabetes is a very common disease that can happen in any age group. It is a serious disease that makes a serious impact on heart, kidneys, nervous system, bloodline and vessels. But mining the data of diabetes patient in an efficient manner is a critical issue. The author had collected the data from various patients either having diabetes or diabetes free. For data mining procedure modified J48 so its accuracy rate can be increased. Author used MATLAB for performing the simulation work with weka as an API to extract the various experimental results [10].

In [7], predictive model of soil fertility has been explained in different steps. In this paper A technique of the decision trees algorithm in data mining is used to predict the fertility of soil followed by performance tuning of J48 decision tree algorithm with the help of meta-techniques such as attribute selection and boosting.

For classification of data and products, the technique of decision tree in used to get valuable results. And these results can be used for analysis and future prediction. In [8] paper the author made an objective to present the enhanced decision tree algorithm that classifies the data. In his work ID3, J48, NBTree are used as the tree classifiers. Then the comparative analysis is done on the basis of parameters like efficiency and performance new enhanced decision tree algorithm (NEDTA).

Table 1. Definition and Procedure of Existing Data Mining Algorithms

Algorithms	Definition and procedure
J48	J48 target variable prediction rules are formed by the algorithm. • Uses top down approach by divide and conquer technique and form tree. • Test attributes are selected by some measures and divide and conquer is applied until no sample leaf is left

ID3	One of the decision tree algorithms called iterative dichotomiser. • Uses greedy approach for creating tree by top down approach. • Selection of attributes which classifies the data at its best on each node and keeps on following this at every node till the tree is not formed is done.
NBTree	Naïve Baseyian classication and decision tree algorithm learning together forms NBtree. • Each node is selected and naïve baseyian algorithm is applied, which classifies the instances. • The naïve baseyian tree is constructed for each leaf.

3. Experimental Setup

There are various numbers of approaches that works on asthma patients but are not capable enough to overcome the various problems of data missing values and classification problems. So in our proposed approach we try to integrate the data pre-processing approach and the classification approach to build a powerful setup for data mining on the asthma patients. In order to carry out the experiment of data mining on a dataset of asthma patients, we have derived the data from various health resources that provide the data on the basis of the surveys they had made. In our data set the numbers of attributes used are 11 and the total numbers of instances are 1776 of different asthmatic and non-asthmatic patients.

Table 2. Various Health Attributes with Description and Domain Used for Building Asthma Dataset

S.No.	Attribute	Description
1	Age	Age of the patients in years.
2	Gender	Gender of the patient whether Male or Female(M/F) (0/1)
3	Begin	Start of the patient record, according to the hospital records
4	End	Ending of the patient record, according to the hospital records
5	Current wheeze	Level of the wheezes (%)
6	Symptoms of severe wheezes	Level of severe wheezes after the current wheezes situation (%)
7	Alcohol consumption	Whether a patient consume alcohol or not(scales 1-5)
8	Drug intake	Whether a person is on any drugs or not (Y/N)(0/1)
9	Physical activity status	Whether a person is physically active or not(Y/N) (0/1)
10	Smoking habits	Whether a person smokes or not(scales 1-5)
11	Hereditary	Whether a person has any previous family asthmatic symptoms or not(Y/N)(0/1)

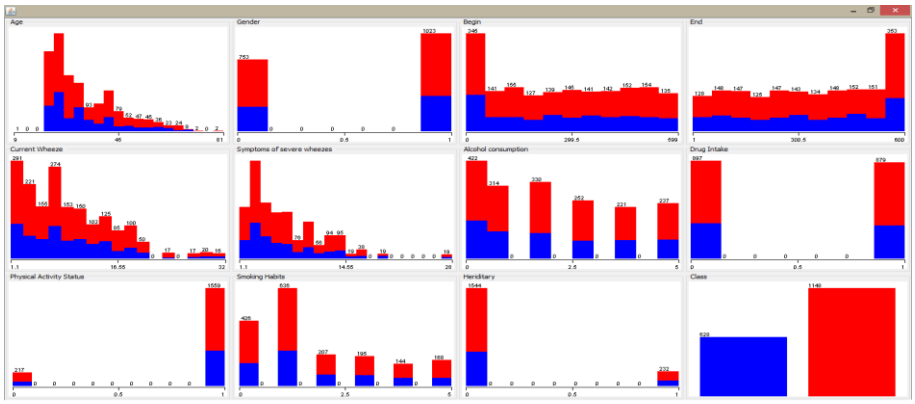

Figure 1. Attribute Based Graphical Representation of Mined Data after Data Pre-Processing Task

Data pre-processing: When a data is uploaded to the server there are many times the data having missing values in them, which can result into false assumptions of various asthma patients. Therefore the missing values need to be taken care of for which we try to introduce the missing value algorithm for data pre-processing process known as normalization. Normalization is the process in which the given dataset is normalized apart from the class attribute in normalized intervals of time. Although in our experimental setup there are no missing values among 1776 attributes of dataset, but still when we apply data pre-processing algorithm the various results come that very intelligently sets the data into various graphical representations .Figure 1 shows the detection of various asthmatic and non-asthmatic patients based on different instances. The blue line symbolizes the asthmatic patient and the red one the non-asthmatic patient

In each graph the combination of red and blue section tells the count of patients, those are asthmatic on the basis of the single attribute. Therefore, every graph shows its own value due to different attribute value in it. Data pre-processing allows the users to pre-process the data before classification which is done above. According to these graphical representations of various attributes and instances, following table can be derived which further mines the data intelligently.

Data classification: The major aspect of the proposed approach is the data classification technique which classifies the data through data classification algorithm into various different classes makes it easy to extract various useful information out of it. The classification algorithm that we are using here is the Random Tree classification algorithm. The proposed algorithm is based on constructing the tree of random amount of values of the instances.

Table 3. Mathematical Values Based Result Derived after Data Mining

Attribute	Minimum value	Maximum value	Mean	StdDev	Distinct values
Age	9	81	33.214	11.705	54
Gender	0	1	0.576	0.494	2
Begin	0	599	266.207	189.578	546
End	1	600	340.15	188.133	56
Current wheeze	1.1	32	9.512	6.497	75
Symptoms of severe wheezes	1.1	28	6.806	4.673	65
Alcohol consumption	0	5	2.139	1.718	6

Drug intake	0	1	0.495	0.5	2
Physical activity status	0	1	0.878	0.328	2
Smoking habits	0	5	1.718	1.589	6
Hereditary	0	1	0.131	0.337	2

The random tree algorithm is tested on the data set of 1776 instances which preform much better than the other classification algorithms. The performance result of the random tree algorithm is 100%, which is the highest among all other classification algorithms. The confusion matrix of the algorithm tells us about the performance and classification capability of the random tree.

Table 4. Confusion Matrix of the Amount of Instances Correctly and Incorrectly Classified

a	B	←Classified as
628	0	a=Asthmatic
0	1148	b=Non-Asthmatic

The confusion matrix above of random tree classification algorithm tells that out of 628 values of 'a' *i.e.* asthmatic patients 0 are the wrongly classified values and out of 1148 values of 'b' *i.e.* non-asthmatic patients 0 are the wrongly classified instances, therefore the performance of the random tree algorithm in quite high. The accuracy of the algorithm is not only counted with the performance percentage, but also with various other factors too needs to be counted and out of which the cost/benefit values and threshold values play a vital role. To be an effective algorithm the two values must be inversely proportional to each other that means the threshold value is higher than the cost value which must be low enough to accommodate the whole values. In figure 2 a graphical comparison is being done with the threshold values and cost/benefit values by comparing a threshold cure and cost/benefit cure with each other which Cleary shows the relation between them. The curves indicate that among the total of 1776 (100%) values of the datasets 1148(64.64%) are asthmatic and 628(35.36%) are considered to be non-asthmatic patients.

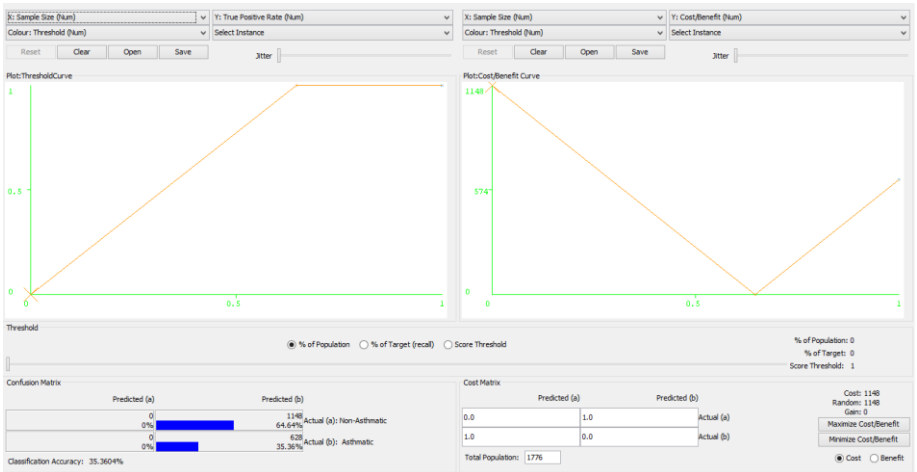

Figure 2. Comparison between Threshold Curve and Cost/Benefit Cure of Asthmatic and Non-Asthmatic Patients

Figure 3. Graphical View of Number of Total Patients Classified as Asthmatic and Non-Asthmatic

To calculate the performance, quality of the algorithm, the algorithm is counted on various QOS parameters that make a comparative analysis with various other existing classification algorithms to make the proposed approach as superior among all of them.

Table 5. QOS Based Comparative Analysis of Various Data Classification Algorithms With The Proposed Approach

Algorithms	Correctly Classified Instances %	Incorrectly Classified Instances %	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (in%)	Root relative squared error(in %)	Total Number of Instance	Time to build
AD Tree	66.3851	33.6149	0.1196	0.4639	0.4732	101.4724	98.9847	1776	0.11
BF tree	64.6396	35.3604	0	0.4571	0.4781	99.9895	100	1776	0.93
Decision Stump	64.6959	35.3041	0.0021	0.4556	0.4773	99.6622	99.8362	1776	0.03
J48	69.1441	30.8559	0.1837	0.4114	0.4535	89.9855	94.8657	1776	0.07
J48graft	69.1441	30.8559	0.1837	0. 4114	0.4535	89.9855	94.8657	1776	0.15
LAD Tree	66.1599	33.8401	0.0669	0.44	0.4678	96.234	97.8579	1776	0.19
NB Tree	64.6396	35.3604	0	0.4572	0.4781	100	100	1776	0.18
Random Forest	98.9302	1.0698	0.9766	0.1684	0.2135	36.8441	44.656	1776	0.21
Random Tree	100	0	1	0	0	0	0	1776	0.02
REP Tree	69.1441	30.8559	0.1949	0.4085	0.4519	89.3543	94.5284	1776	0.07
Simple Cart	64.6396	35.3604	0	0.4571	0.4781	99.9895	100	1776	0.05

4. Conclusion

In this paper, we have proposed an integrated approach for mining the data of asthma patients for which normalize approach is used for data pre-processing and random tree approach for data classification. We have made an experimental setup for our algorithm to execute and different performance graphs and data mined values have been taken after data pre-processing and classification of data. At the end we made a comparative analysis of random tree algorithm with many other classification approaches and conclude that the random tree algorithm is the only algorithm whose capability of classifying the instances correctly is 100% without any wrongly classified instances and zero error rates. The CPU utilization is also very low as the time taken to mine the whole data set is very less *i.e.* 0.02 seconds which is least among all algorithms.

References

[1] A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester and P. Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing", Journal of Biomedical Informatics, vol. 43, **(2010)**, pp. 342–353.
[2] H. Xia, I. Asif and X. Zhao, "Cloud-ECG for real time ECG monitoring and analysis", computer methods and programs in biomedicine, vol. 110, **(2013)**, pp. 253-259.
[3] J. W. Dexheimer, T. J. Abramo, D. H. Arnold, M. P. H. Kevin Johnson, M. S Yu Shyr Fei Ye Kang-Hsien Fan Neal Patel and M. S. Dominik Aronsky, "Implementation and Evaluation of an Integrated Computerized Asthma Management System in a Pediatric Emergency Department: A Randomized Clinical Trial", International Journal of Medical Informatics.
[4] S. Pandeya, W. Voorsluys a, S. Niua, A. Khandokerb and R. Buyyaa, "An autonomic cloud environment for hosting ECG data analysis services", Future Generation Computer Systems, vol. 28, **(2012)**, pp. 147–154.
[5] V. Vaithiyanathan1, K. Rajeswari2, K. Tajane3 and R. Pitale, "Comparison Of Different Classification Techniques Using Different Datasets", International Journal of Advances in Engineering & Technology, May **(2013)**.
[6] A. Goyal and R. Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", International Journal of Applied Engineering Research, ISSN 0973-4562.
[7] J. Gholap, "Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility".
[8] H. Kaur and H. Kaur, "Classification of data using New Enhanced Decision Tree Algorithm", International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS).
[9] T. R. Patil and Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal Of Computer Science And Applications, ISSN: 0974-1011, vol. 6, no.2, April **(2013)**.
[10] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887), vol. 98, no. 22, July **(2014)**.
[11] D. Oletic, "Wireless sensor networks in monitoring of asthma", International convention micro, vol. 34, **(2011)**.
[12] G. K. Nee, M. A. Syafiq, S. K. Sugathan, C. Y. Yie and E. A. P. Akhir, "The Development of a Rule-based Asthma System" Information Technology (ITSim)", 2010 International Symposium in Date 15-17, June **(2010)**.
[13] M. A. khassaweneh1, S. B. Mustafa1 and F. A. Ekteish2, "Asthma Attack Monitoring and Diagnosis: A Proposed System", Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on Date17-19, December **(2012)**.
[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing and Management, vol. 45, **(2009)**, pp. 427–437.
[15] S. Rietveld, M. Oud and E. H. Dooijes, "Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural network", Computers and Biomedical Research, vol. 32, **(1999)**, pp. 440–448.
[16] R. Bozorgmanesh a, M. Otadi b, A. A. Safe Kordi c, F. Zabihi d and M. B. Ahmadi, "Lagrange Two-Dimensional Interpolation Method for Modeling Nanoparticle Formation During RESS Process", Int. J. Industrial Mathematics, vol. 1, no. 2, **(2009)**, pp. 175-181.
[17] C. E. Sabel a, W. Kihal b, D. Bard and C. Weber, "Creation of synthetic homogeneous neighborhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France", Social Science & Medicine.
[18] T. Pham and M. Wagner, "Ambiguity reduction in speaker identification by the relaxation labeling process, Pattern Recognition, vol. 32, no. 7, **(1999)**, pp. 1249–1254.

[19] A. Rosenthal, P, Mork, M. H. Li, J. Stanford, D. Koester and P. Reynolds, "Cloud computing: A new business paradigm for biomedical information sharing", Journal of Biomedical Informatics, vol. 43, **(2010)**, pp. 342–353.
[20] H. Xia, I. Asif and X. Zhao, "Cloud-ECG for real time ECG monitoring and analysis", computer methods and programs in biomedicine, vol. 110, **(2013)**, pp. 253-259.
[21] J. W. Dexheimer, T. J. Abramo, D. H. Arnold, M. P. H. Kevin Johnson, M. S. Y. Shyr Fei Ye Kang-Hsien Fan Neal Patel, M. S. Dominik Aronsky, "Implementation and Evaluation of an Integrated Computerized Asthma Management System in a Pediatric Emergency Department: A Randomized Clinical Trial", International Journal of Medical Informatics.
[22] S. Pandeya, W. Voorsluys a, S. Niua, A. Khandokerb and R. Buyyaa, "An autonomic cloud environment for hosting ECG data analysis services", Future Generation Computer Systems, vol. 28, **(2012)**, pp. 147–154.
[23] V. Vaithiyanathan1, K. Rajeswari2, K. Tajane3 and R. Pitale, "Comparison Of Different Classification Techniques Using Different Datasets", International Journal of Advances in Engineering & Technology, May **(2013)**.
[24] A. Goyal and R. Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms", International Journal of Applied Engineering Research, ISSN 0973-4562.
[25] J. Gholap, "Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility".
[26] H. Kaur and H. Kaur, "Classification of data using New Enhanced Decision Tree Algorithm", International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS).
[27] T. R. Patil and Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science And Applications, ISSN: 0974-1011, vol. 6, no.2, April **(2013)**.
[28] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887), vol. 98, no. 22, July **(2014)**.
[29] D. Oletic" Wireless sensor networks in monitoring of asthma", International convention micro, vol. 34, **(2011)**.
[30] G. K. Nee, M. A. Syafiq, S. K. Sugathan, C. Y. Yie and E. A. P. Akhir, "The Development of a Rule-based Asthma System", Information Technology (ITSim), 2010 International Symposium in Date 15-17, June **(2010)**.
[31] M. A. khassaweneh1, S. B. Mustafa1 and F.A. Ekteish2, "Asthma Attack Monitoring and Diagnosis: A Proposed System", Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on Date 17-19, December **(2012)**.
[32] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing and Management, vol. 45, **(2009)**, pp. 427–437.
[33] S. Rietveld, M. Oud and E. H. Dooijes, "Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural network", Computers and Biomedical Research, vol. 32, **(1999)**, pp. 440–448.
[34] R. Bozorgmanesh a_, M. Otadi b, A. A. Safe Kordi c, F. Zabihi d and M. B. Ahmadi, "Lagrange Two-Dimensional Interpolation Method for Modeling Nanoparticle Formation During RESS Process", Int. J. Industrial Mathematics, vol. 1, no. 2, **(2009)**, pp. 175-181.
[35] C. E. Sabel a, W. Kihal b, D. Bard and C. Weber, "Creation of synthetic homogeneous neighborhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France", Social Science & Medicine.
[36] T. Pham and M. Wagner, "Ambiguity reduction in speaker identification by the relaxation labeling process", Pattern Recognition, vol. 32, no. 7, **(1999)**, pp. 1249–1254.

Author

Abhinav Hans, was born on 15-09-1990.He completed his B. Tech in Computer science and engineering from Lovely Professional University, Phagwara, Punjab, India in the year of 2013.He is pursuing his M. Tech in Computer Science and Engineering from Guru Nanak Dev. University, R C, Jalandhar. His field of interest are cloud computing, biomedical, big data, wireless sensor network, body area network. Till now he has various number of publications out of which much of them are in IEEEXPLORE and International Journals.