IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Physiological Variability Analysis Using Pre-existing Data Mining Techniques

Shagun Goyal[1], Sunila Godara[2]

PG Student, Computer Science Department, Guru Jambheshwar University S& T, Hisar, Haryana[1]

Assistant Professor, Computer Science Department, Guru Jambheshwar University S& T, Hisar, Haryana [2]

Abstract: Background/Objective: To analyse the efficiency and performance of pre-existing data mining techniques for development of novel predictor model for mining the response shown by physiological parameters. The medicines here being tested are physiological variability affecting medicines. Methods/Statistical Analysis: The data has been have been analysed using WEKA (version 3.7) tool for the following techniques – Classification Via Regression, Randomized filtered classifier, IBk and RandomForest technique. Findings: Usage pre-existing data mining techniques for development of novel prediction models for mining physiological variability data. An overall comparison of various techniques has also been made on the basis of various performance parameters like sensitivity, specificity, precision & F-measure. It has been found that Randomized Filtered classifier is the best suitable technique amongst all the techniques for such a use. Applications: it can be used for mining the physiological variability responses, thereby helping to check the effectiveness of medicines.

Keywords: Data Mining, Physiological Variability, ClassificationViaRegression, Randomized Filtered Classifier, Random tree classifier, IBk technique.

I. INTRODUCTION

Data mining is a computational process of extracting useful information i.e. knowledge from raw, unprocessed data sets which are available in enormous amounts in scientific, medical, demographic, financial ,marketing and many other fields[8]. Classification and prediction techniques which are a major sub-part of data mining techniques.

In medical fields, the physiological variability (PV) is of great significance as it helps in detection of diseases without invasively intervening with the body. Suppose a person is walking on the road and suddenly a snake crosses by. His heartbeat may increase suddenly and once the snake passes; it turns to normal again. These subtle changes take place in the body at much lower level, even when we are in the state of rest. The changes brought by in our body by these activities when we are in state of rest termed as Physiological Variability [10].

Two types of activities that take place in our body which comes under autonomic nervous system (ANS), are sympathetic and para-sympathetic changes which affect the working of internal organs. These functions are involuntary and regulates body functions such as heart rate, digestion etc. Sympathetic and para-sympathetic nervous systems have opposite effect on a particular system i.e. while Sympathetic system increases the activity of an organ; the Para-sympathetic would have opposite effect on the same. It is necessary to balance or nullify the effect of sympathetic changes by moderating them by contrasting effects of para-sympathetic changes. That's why para-sympathetic system should be kept in continuous force to maintain the order of body smooly.

With the advent of large number of non-invasive data mining techniques, detecting of physiological parameters have become quite easy. Out of these we can mine various PV parameter values which help to detect health condition of an individual. Also, the effect of certain medicines can be asserted by mining the responses to see the effect of physiological variability affecting medicines by mining the reading taken before and after medicine intake. Our target is to mine such minor physiological responses from the data collected using pre-existing data mining techniques which are due to intake of medicines. There is large amount of noise is also present in the body which may get detected that's why this task of detecting interventional changes due to medicine is crucial and difficult task.

II. BACKGROUND

The roots of physiological variability detection were found in traditional Indian Medical System (AYURVEDA) thousand years ago. They have formulated a subjective method called 'Nadi Pariksha' to measure the health index of an individual. In this method, the Ayurvedic physicians palpated with three fingers, at the wrist location of the patient, the area above the radial artery and intellectually observed the pulse propagation, rhythm, and pulse pressure, pulse volume etc. to attain the diagnosis of the patient. The outcome depended on the knowledge and proficiency of the physician and thus was subjectively biased [12].

Today, there are considerable numbers of invasive and non-invasive techniques, with the help of which physiological variability can be easily detected. As today it is possible to

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

find the physiological variability parameters of a person easily using non-invasive methods, large amount of data has been accumulated and has been stored in databases. Use of Peripheral Pulse Analyzer(PPA) is also a non-invasive way of detecting the same.
Physiological variability is basically determined by three major factors: - Heart Rate Variability (HRV), Peripheral Blood Flow (PBFRV) & Morphological Index(MI)[2]. PPA determines 33 parameters related to HRV, PBFRV and MI [11].

PPA: Peripheral Pulse Analyzer (PPA) is a computer based system used for the study of physiological variability. Its distinctive feature is that it yields heart rate variability, peripheral blood flow variability and morphology index variability from single data acquisition session from the patient [5]. PC controls the data acquisition. It is serially connected to the acquisition unit. The PC performs the variability analysis and transfers the same to the database management system.

Classifier Techniques being used for classification:

A. Classification Via Regression
This technique comes under meta classifier algorithms which are basically used in areas of predictive data mining. It combines the predictions from multiple models and is of great use where type of models used in project is extremely different. Regression is a technique in which an equation is found to fit the formula. In multiple regression, more than one input is used and helps to build complex equation like quadratic equations. Using regression in classification helps to improve the accuracy. Here class is binarized and one for each class value one regression model is built. This is a recent successful technique used to predict continuous numeric values, henceforth is suitable for datset.

B. Randomized Filter classifier
Randomized Filter classifiers are also a part of meta classifier algorithms. Meta classifiers are amalgamation of more than one classifier. The training and test data determines the structure of the classifier as well as filter. The label class is passed through an arbitrary filter, which with the help of this training data builds its structure and then test data instances are also processed on the same structure.

C. Random forest
Random forest learning technique is a part of tree algorithms. It performs regression and classification by constructing multitude of decision trees at the training time. Given by Leo Brieman, this framework builds a forest of un-correlated trees using Classification and Regression trees (CART) combined with randomized node optimization and bagging[i][14].

D. IBk (K stands for number of neighbors)
This technique is a part of lazy learning algorithms. It is similar to K-nearest neighbour classifier (KNN classifier). It selects appropriate value of k based on cross-validation. It can also do distance weighing.

III. INFORMATION ABOUT DATA SET

The data has been collected using Peripheral Pulse Analyzer. The data-set used for deducing results was taken from Electronics Division, Bhabha Atomic Research Centre, Mumbai(BARC). Few key points of the data set are explained as follows: - each record comprises of values equipped from a single reading by PPA where each attribute is a parameter value determined by PPA. This data contains 4050 records & each record comprises of 33 attributes. Each medicine has been tested on a subgroup of 15 persons. 30 readings are taken for each person, of which 25 are taken before medicine and 5 are taken after giving the medicine. Each reading has been taken at an interval of 5 minutes. Data has been recorded preferably on three successive days for each person. On each of the days 5 readings were noted before as well as after the placebo/medicine. Placebo was administered on first two days and has been referred as placebo and anxiety. In the similar manner, Medicinal effect was administered on third day. Thus subjects in one subgroup (15) * readings for one person(30)=450 readings comprises a single subgroup i.e. for a single medicine. 450 * 9 subgroups=4050 readings (the total readings as given above)

Medicinal Response: - Conventionally, the mining of such response was done in order to get class labels for test data as follows: - the responses of the medicines were analysed graphically. When the graph was plotted for each person, for each parameter against the readings taken on each day; the readings taken on the day 3rd after medicine intake, was checked with other plotted before medicine intake readings. If the graph peak showing parameter value after having medicine significantly high or low, then it was termed as response. This work was done manually using conventional graphical methods by BARC, Mumbai. An example of the significant response is for one of the parameters is:

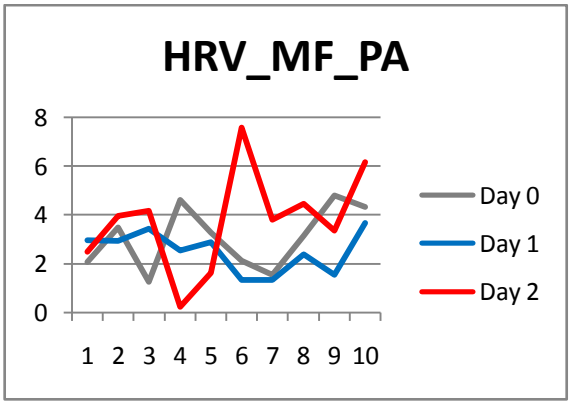

Fig 1: Significant response

In the above graph, the grey line represents anxiety readings taken on day 0, blue line represents placebo readings and day3 readings are represented by red line.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

IV. PROPOSED FRAMEWORK

The following steps hav
e been taken in order to mine the data effectively:-
We have tried to automate this task of analysis using data mining techniques. As the data is a mere collection of readings where a machine cannot distinguish the readings taken before medicine and the one taken after medicine nor the subject(person) number from which another medicine testing has been started ; the machine has to be made to learn this. This can be done by pre-coding it manually as per the need of segregation. Thus the data has been reduced to manageable format and then have been given for analysis. The proposed framework for analyzing such data is as follows:-

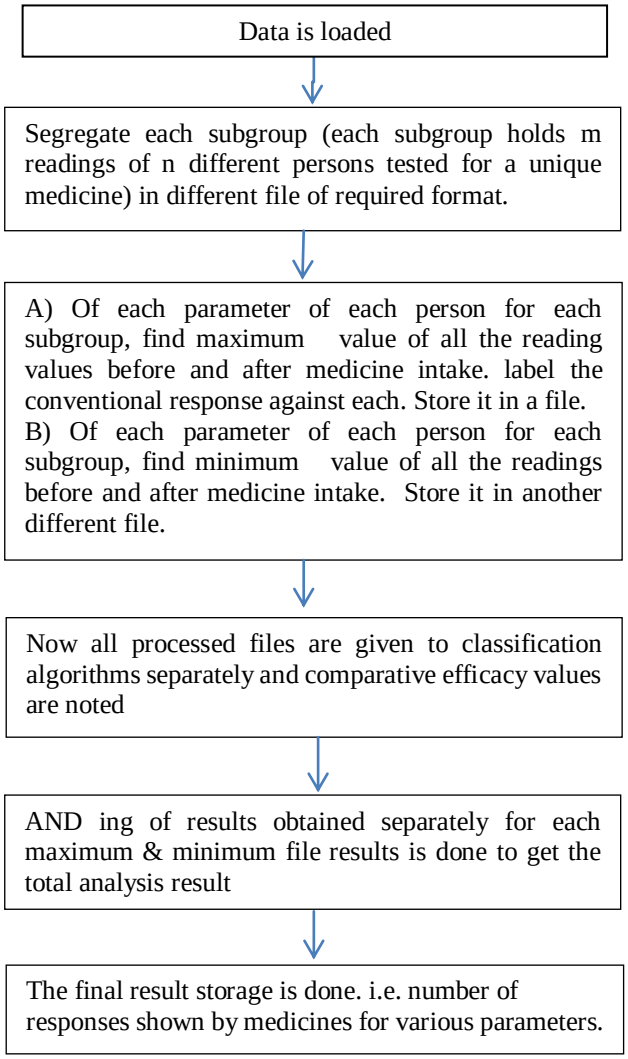

Fig 3: Proposed Methodology

Here 9 different medicines' response analysis has been performed using 10 fold cross validation. The **average** of the performance metrics given by 9 different medicines for various measures i.e (Sensitivity, specificity, accuracy, precision) is presented in result section. They depicts the worth of the algorithms this data.

Processed file now contains three parameters:-

Table-II For maximum response file

S. No.	parameter	Description
1.	Max_val_ before med. reads	Maximum (value) of all the readings taken before a particular parameter for a particular person
2.	Max_val_ after med. reads	Maximum (value) of all the readings taken after medicine intake for a particular parameter of a particular person
3.	Response (class label)	Conventional response (yes /no) [It was conventionally mined using graphical methods]

Table-III For minimum response file

S.No.	Parameter	Description
1.	Min_val_ before med. Reads	Minimum (value) of all the readings taken before a particular parameter for a particular person
2.	Min_val_ after med. Reads	Minimum value of all the readings taken after medicine intake for a particular parameter of a particular person
3.	Response (class label)	Conventional response (yes /no) [It was conventionally mined using graphical methods]

V. RESULTS

For each of the methods, the data has been tested on 9 different subgroups separately i.e is for 9 different medicines to fetch highly precise efficiency measures. Results have been combined for minimum and maximum response for each subgroup. An average of all efficiency measures given by various subgroups has been shown.
Before we actually study the efficacy measures, following terms should be kept in mind:

a) True Positives (TP):- It refers to number of tuples that were predicted true and actually found to be true.

b) Truly Negatives (TN):- It refers to number of tuples that were predicted false and were actually false.

c) False Positives (FP):- It refers to number of tuples that were predicted true and were actually false.

d) False Negatives (FN):- It refers to number of tuples that were predicted false and were actually to be true.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Accuracy/ Overall Recognition Rate: The percentage of test set tuples that are correctly classified by a classifier on a given test set is termed as Accuracy [2].

Accuracy = TP + TN / (TP + FP + TN + FN)

It reflects how aptly a classifier recognizes tuples of various classes. The classifiers trained here are to classify data tuples as "response" or "no response". The accuracy rate of 97% appears to be accurate, but when conventionally measured, it has been found that among the entire medical data only 5-6% actually shows positive response.

Undoubtedly, An accuracy rate of 97% or above is not acceptable as this implies that above algorithms are labelling maximum of "no response" tuples correctly[9].

Thus we have to check the other measures in order to determine the efficacy of these algorithms in mining this data.

Besides the other measures, the most important measures which validates the efficacy of the algorithm sufficiently are:-

Sensitivity/Recall/True Positive Rate:- It is defined as the ratio of positive tuples that are correctly identified to all the tuples that are identified as positive.

Sensitivity=TP/ (TP +FN)

Here the highest sensitivity of Randomized Filter classifier shows that this is the best learning algorithm as it detects the maximum number of positive responses.

Specificity/True Negative Rate:- It is defined as the ratio of negative tuples that are correctly identified to the total tuples identified as negative. Here, Specificity insinuates the percentage of negative responses which are correctly identified [6].

Specificity=TN/ (TN+FP)

Table-II Efficacy parameters of various algorithm

S.No.	Algorithm	Sensitivity	Specificity	F-measure	Accuracy	Precision
1	Classification_Via_Regression	21.286%	99.884%	83.100%	97.465%	33.891%
2	Randomized Filtered Classifier	51.000%	99.350%	70.893%	98.114%	59.323%
3	Random Forest Technique	37.000%	99.080%	80.204%	97.867%	50.639%
4	IBk Technique	44.000%	99.631%	65.993%	98.172%	52.798%

The extremely high specificity of all the algorithms is because of high presence of conventionally high number of negative responses.

Precision: - It is defined as ratio of tuples that are correctly recognized as positive to all the tuples that are recognized as positive. In other words, Precision signifies the ratio of tuples correctly identified as positive response that actually are positive.

Precision = TP/ (TP + FP)

It is understandable that the specificity and accuracy are prone to be high as they are consequents of number of true negatives. Number of negatives constitutes 95% data approximately, which are very large. In case of specificity and precision, numbers of true positives tuples are main determining factors. These tuples are very present in very less number, hence the efficiency of techniques determining such responses are varied and are imperative for this research.

We can see that Randomized Filter Classifier technique has highest sensitivity amongst all which implies that of all the positively detected tuples, true positives are highest in this technique. Besides this, this technique has precision of 70.893%, which implies 70.893% tuples will be correctly identified and 29.107% tuples are falsely identified as positive. Hence, comparing to all the techniques this one is most suitable. ClassificationViaRegression technique holds the highest precision value of 82.1% but delimits because of low sensitivity of 21.286%.

F-measure/ F- score:- It is the harmonic mean of precision and recall. A high F-score value implies better efficacy of the algorithm.

F-measure= 2.Precision.Recall / (Precision + Recall)

We can see clearly from the following bar chart that the Randomized Filter Classifier gives the highest performance, as it has all of it's performance parameters of higher value, followed by IBk technique, Random forest Classification and Classification via Regression. The reader should not miss that authors have taken the target of mining the response. Such response may be proving or curing as the type of medicines being taken are homeopathic medicines. Both type of responses, which are significantly higher as well as lower are taken into account.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

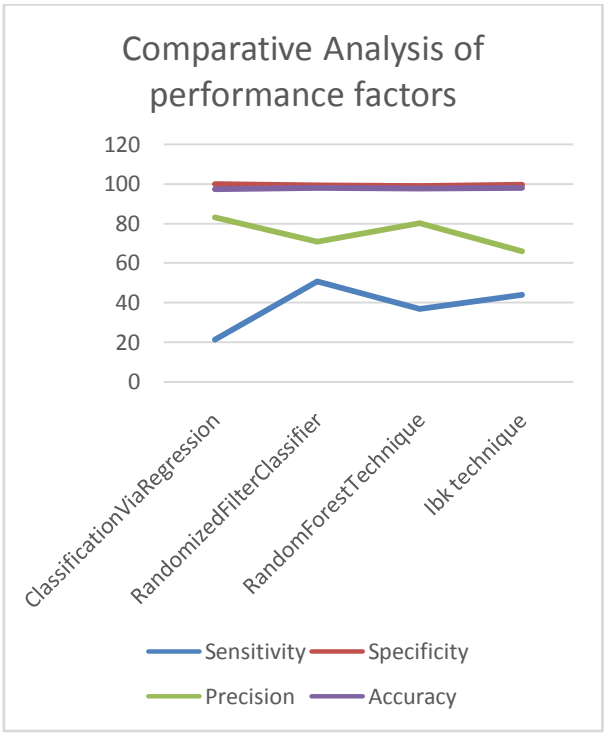

Fig. 2 A comparative analysis of the above techniques

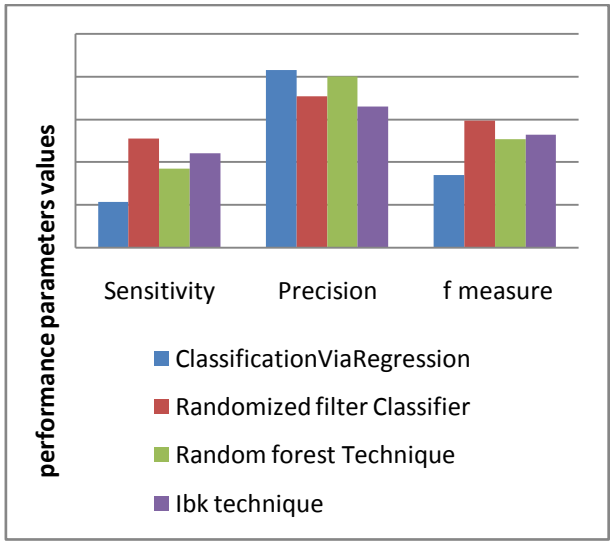

Fig. 3 A comparative analysis of the above techniques

VI. CONCLUSION

In this paper, we have seen the performances of various algorithms on such specific data for interventional analysis of physiological variability. Randomized filter classifier is the best algorithm we have found to analyse this data as. Besides the best trials we say that there are always chances of improvement. Medical responses should be mined as precisely as possible. They may help to define the best suitable medicine. We also anticipate that a new data mining algorithm which fits well for such data can be written which can always be used for mining data. Besides this, other data mining learning techniques may also be researched for better performance.

ACKNOWLEDGMENT

With the kind blessings of god and support of various highest minds, the authors have made this research worthy. The authors express their profound thanks to Bhabha Atomic Research Centre, Mumbai. Also we are indebted to **Dr. G.D Jindal,** Department of Bio-medical Engineering and Technology Navi Mumbai, Maharashtra for helping in algorithm development, **Sh. R.K Jain**, Scientific Officer (H+), Bhabha Atomic Research Centre, Mumbai, Maharashtra & **Ms. S.N Bhat,** Senior Research Fellow (BRNS), Electronics Division, BARC, Mumbai who have manually mined the data and provided us with class labels required to check efficiency of the data mining algorithms. We also thank **Mr. Sahil Kakkar,** PG student, Guru Jambheshwar University S & T, Hisar, Haryana for instilling the innovative ideas for the proposed methodology.

REFERENCES

[1] Sunila Godara and Shagun Goyal, "Automation of Interventional Physiological Variability Analysis: A Review", International Journal of Advanced Research in computer and Communication Engineering, vol 5, Issue 4, 2016.
[2] S. Godara and R. Singh, "Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis", Indian Journal of Science and Technology, vol. 910, 2016.
[3] P. Melillo, R. Izzo, A. Orrico, P. Scala, M. Attanasio, M. Mirra, N. De Luca and L. Pecchia, "Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis", PLOS ONE, vol. 10, no. 3, p. e0118504, 2015.
[4] Saha and D. Nandi, "Data Classification based on Decision Tree, Rule Generation, Bayes and Statistical Methods: An Empirical Comparison", International Journal of Computer Applications, vol. 129, no. 7, pp. 36-41, 2015.
[5] G. Jindal, R. Jain, V. Sinha, S. Mandalik, P.Tanawade, C. Pithawa, P. Kelkar and A. Deshpande, "Early Detection of Coronary Heart Disease Using Peripheral Pulse Analyzer", BARC newsletter, no. 326, 2014.
[6] M. Tchikviladzé, M. Gilleron, T. Maisonobe, D. Galanaud, P. Laforêt, A. Durr, B. Eymard, F. Mochel, H. Ogier, A. Béhin, T. Stojkovic, B. Degos, I. Gourfinkel-An, F. Sedel, M. Anheim, A. Elbaz, K. Viala, M. Vidailhet, A. Brice, C. Jardel and A. Lombès, "A diagnostic flow chart for POLG- related diseases based on signs sensitivity and specificity", Journal of Neurology, Neurosurgery & Psychiatry, vol. 86, no. 6, pp. 646-654, 2014.
[7] Pooja Mittal, Nasib Singh Gill, "A Comparative Analysis of Classification Techniques on Medical data sets", International Journal of Research in Engineering and Technology, vol. 03, no. 06, pp. 454-460, 2014
[8] P. Drotár and z. Somparative study of machine learning techniques for supervised classification of biomedical data", aei, vol. 14, no. 3, pp. 5-10, 2014
[9] J. Han and M. Kamber, Data mining. Haryana, India: Elsevier, 2012.
[10] N. Mishra, K. Muraleedharan, A. Paranjpe, D. Munta, H. Singh and C. Nayak, "An Exploratory Study on Scientific Investigations in Homeopathy Using Medical Analyzer", The Journal of Alternative and Complementary Medicine, vol. 17, no. 8, pp. 705-710, 2011.
[11] G.D Jindal, K.K Deepak and R.K Jain, A Handbook on Physiological Variability Advanced Applications of Physiological Variability (AAPV-2010). Navi Mumbai: BRNS, DAE, 2010.
[12] Ananthakrishnan T.S., Pithawa C.K., "Introduction to Physiological Variability: in A handbook on physiological variability" In: Jindal GD, Deepak KK, Jain RK, editors. Electronics Division BARC, 1-17, 2010
[13] C. Lin, J. Wang and P. Chung, "Mining Physiological Conditions from Heart Rate Variability Analysis", IEEE Comput. Intell. Mag., vol. 5, no. 1, pp. 50-58, 2010.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

[14] O. Maimon and L. Rokach, Data mining and knowledge discovery handbook. New York: Springer, 2005.
[15] Leo Breiman, "Random Forests", Machine Learning, 2001

BIOGRAPHY

Ms. Shagun Goyal, student of M.Tech final year of computer Science field at Guru Jambheshwar University, Hisar is carrying out her research work in data mining field. She has carried out a a part of her dissertation work at Bhabha Atomic Research Centre, Mumbai. Her area of interest is data mining and software engineering.

Ms Sunila Godara received MSc degree in Computer Science & Engg from Guru Jambheshwar University of Science & Technology, HISAR. She is working as Assistant Professor in Dept. of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, HISAR. She has published more than 25 papers in national and international journals and conferences. Her research areas are Data Mining and Database Management System.