International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-4 Issue-13 December-2013

An Overview on Data Mining Approach on Breast Cancer data

Shiv Shakti Shrivastava[1], Anjali Sant[2], Ramesh Prasad Aharwal[3]

Abstract

This paper gives the current overview of use of data mining techniques on breast cancer data. This paper also gives the study of data mining on medical domain which has already done from researchers. In this paper we use classification data mining techniques on breast cancer data with using data mining software. A huge amount of medical records are stored in databases. Data are produce from different sources and continuously stored in depositories. These databases are more complicated for the point of analysis. Data Mining is a relatively new field of research whose major objective is to acquire knowledge from large amounts of data.

Keywords

Data Mining, Breast Cancer Data, Weka, Decision Tree.

1. Introduction

Breast cancer is the second mainly common disease in women in India and this disease is increasing annually. The lack of awareness initiatives, structured viewing, and affordable treatment facilities continue to result in poor survival. Cancer of the breast is the second most common human neoplasm, accounting for around one quarter of all cancers in females after cervical carcinoma. We present the overview of research in use of data mining techniques in breast cancer. Data mining has become a popular technology in current research and for medical domain applications [29, 17]. Breast cancer has become the leading cause of death in women in developed countries. The rest of the paper is organized in two parts. First part concern with review of research in the application of data mining. techniques on breast cancer and second part consist experimental work. First part is organized as follows: The next section discusses About Knowledge discovery and database its subsequently discussed data mining and its tasks that are related to data mining techniques. Next section concerns with previous work of research review of application of data mining in breast cancer data. Second part consist dataset description and experiment on weka and discusses the results and future work.

Shiv Shakti Shrivastava, Research Scholar Mewar University, Chittorgarh (Raj.)
Anjali Sant, Professor, BITS, Bhopal
Ramesh Prasad Aharwal, Asstt. Prof., Department of Mathematics and Computer Application Govt. P. G. College Damoh (M.P.) India

2. Knowledge Discovery and Database (KDD)

The Definition of KDD given by **Fayyad, et al.[**13] " KDD is the non trival process of identifying, novel, Potentially useful and ultimately understandable patterns in data ". **Data Mining is the crucial step of KDD which** automatically searching large volumes of data for patterns using algorithms such as classification, clustering, etc. The general difference between "KDD and Data mining is that overall process of discovering useful knowledge from data while data mining refers to the application of algorithms for extracting patterns from data without the additional steps of KDD process " [13].

2.1Data Mining

Data mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data. According to this definition data mining is the step that is responsible for the actual knowledge discovery [29]. To emphasize the necessity that data mining algorithms need to process large amounts of data, the desired patterns has to be found under acceptable computational efficiency limitations. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment.[K. Rajesh, V. Sangeetha (2012)[23]]. Data mining has widely use in various do mains such as medical, healthcare, higher education, telecommunication etc.

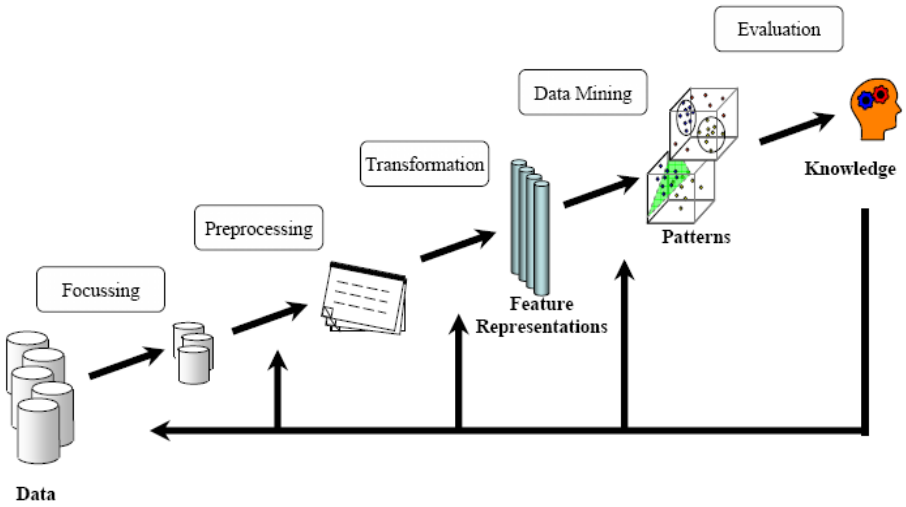

Figure 1 KDD Process

In 1991 it was alleged that the amount of data stored in the world doubles every twenty months. At the same time, there is a growing realization and expectation that data, intelligently analyzed and presented, will be a valuable resource to be used for a competitive advantage[18]. To cross the growing gap between data generation and data understanding, there is an urgent need for new computational theories and tools to assist humans in extracting useful knowledge from the huge volumes of data. These theories and tools are the subject of the emerging field of Knowledge Discovery in Databases (KDD), or Data Mining (DM), Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses [24]. Three steps involved are:

- Exploration
- Pattern identification
- Deployment

2.2Datamining Task

The main tasks of data mining involve extracting meaningful new information from the data. Knowledge discovery comes in two flavors:

•Predictive: Classification, Regression, Time series analysis, Prediction

•Descriptive: Clustering, Summarization, Association rules, Sequence discovery

2.3Decision Tree

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Decision trees are powerful classification algorithms that are becoming more and more popular with the growth of data mining in the field of information systems. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 [21,22], and Breiman et al.'s CART [5]. Decision trees can be used for segmentation of the original dataset (each segment would be one of the leaves of the tree) like segmentation of customers, products, and sales regions etc so that similar data can be grouped in one segment. The segmentation is done for the prediction of some important piece of information [11].

3. Previous Study on Application of Data Mining In Breast Cancer

This section consists of the review of various research papers and review articles on data mining techniques applied in breast cancer dataset. The various common data mining methods and techniques used for breast cancer diagnosis are Mammography, Biopsy, Positron Emission Tomography and Magnetic Resonance Imaging. The results obtained from these methods are used to recognize the patterns which are aiming to help the doctors for classifying the malignant and benign cases. There are various data mining techniques and machine learning algorithms that are

applied to analyze and find some decision rules in breast cancer data.

Delen D, Walker G, Kadam A. (2005)[10] In their work, they have done the study of as the starting point of our research. In his study, preprocessed the SEER data with 433,272 records for breast cancer to remove redundancies and missing information. The resulting data set had 202,932 records, which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659) depending on the Survival Time Recode (STR) field. After that they applied data mining algorithms on these data sets to predict the dependent field from 16 predictor fields. The results of predicting the survivability were in the range of 93% accuracy.

As pointed out by Testard P.Vaillant (2010) [27] "information, dialog and more patient involvement in the decision-making process" are key words in dealing with cancer, therefore a major challenge in the field of medical counseling is to provide physicians and radiologists with adequate tools to help them to assess breast cancer risk of their patients and to show easily how risk factors impact global risk.

In [4] they built a risk prediction model using a logistic regression on the Breast CancerSurveillance Consortium (BCSC) database which contains 2.4 million screenings mammograms and associated self-administered questionnaires. Two logistic regression risk models were constructed with 4 or 10 risk factors depending on the menopausal status. Compared to Gail's model, it gains the use of breast density and hormone therapy. As we will use the same database, it is worth highlighting that reported area under ROC curve (see performance measurement in section IV-D) was 0.631 for premenopausal women and 0.624 for postmenopausal women.

In A. Endo, T. Shibata, and H. Tanaka,(2008)[12] they implemented common machine learning algorithms to predict survival rate of breast cancer patient. This study is based upon data of the SEER program with high rate of positive examples (18.5 %). Since this study aims at classifying examples in two classes, authors did not used ROC curve to assess performances results but accuracy, specificity and sensitivity. Logistic regression had the highest accuracy, artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

In [25], they point out a comparison among the capabilities of various neural networks such as Multilayer Perceptron, Self Organizing Map, Radial Basis Function and Probabilistic Neural Network which are used to classify WBC and NHBCD data. The performance of these neural network structures was investigated for breast cancer diagnosis problem. This work showed that statistical neural networks can be effectively used for breast cancer diagnosis as by applying several neural network structures a diagnostic system was constructed that performed quite well.

Anunciacao Orlando, Bruno C. Gomes, Susana Vinga, et al (2010) [1], they present the applicability of decision trees for detection of high-risk breast cancer groups over the dataset. This dataset was produced by the Department of Genetics of faculty of Medical Sciences of Universidade Nova de Lisboa with 164 controls and 94 cases in WEKA machine learning tool. They found a high-risk breast cancer group composed of 13 cases and only 1 control, with a Fisher Exact Test value of 9.7*10-6 and a p-value of 0.017. These results showed by decision tree.

Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed (2010) [2] they have investigated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases.

Wei-pin Chang, Der-Ming and Liou(2010) In [6] they point out in their study that the genetic algorithm model yielded better results than other data mining models for the analysis of the data of breast cancer patients in terms of the overall accuracy of the patient classification, the expression and complexity of the classification rule. They were used these techniques artificial neural network, decision tree, logistic regression, and genetic algorithm for the comparative studies and the accuracy and positive predictive value of each algorithm were used as the evaluation indicators. They were WBC database incorporated for the data analysis followed by the 10-fold cross-validation. The results showed that the genetic algorithm described in the study was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible[11].

K. Rajiv Gandhi, Marcus Karnan and S. Kannan(2010) [14] In their paper constructed

classification rules using the Particle Swarm Optimization Algorithm for breast cancer datasets. In this study to cope with heavy computational efforts, the problem of feature subset selection as a pre-processing step was used which learns fuzzy rules bases using GA implementing the Pittsburgh approach. It was used to produce a smaller fuzzy rule bases system with higher accura . The resulted datasets after feature selection were used for classification using particle swarm optimization algorithm. The rules developed were with rate of accuracy defining the underlying attributes effectively.

J. Padmavati(2011) [20] they performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression.

Sudhir D. Sawarkar et al(2006) [26] in their study they applied SVM and ANN on the WBC data .The results of SVM and ANN prediction models were found comparatively more accurate than the human being. The 97% high accuracy of these prediction models can be used to take decision to avoid biopsy.

Sepehr M. H. Jamarani et al(2005) [15] they presented an approach for early breast cancer diagnosis by applying combination of ANN and multi wavelet based sub band image decomposition. The proposed approach was tested using the MIAS mammographic databases and images collected from local hospitals. The best performance was achieved by BiGHM2 multi wavelet with areas ranging around 0.96 under ROC curve. The proposed approach could assist the radiologists in mammogram analysis and diagnostic decision making [9].

M. Lundin et al (1999) In [19] has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku to evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. The values of ROC curve for 5 years was evaluated as 0.909,for 10 years 0.086 and for 15 years 0.883, these values were used as a measure of accuracy of the prediction model. They compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy.

Chih-Lin Chi et al(2007)[8] they used the Street's ANN model for Breast Cancer Prognosis on WPBC data and Love data. In their research they used recurrence at five years as a cut point to define the level of risk. The applied models successfully predicted recurrence probability and separated patients with good (>5 yrs) and bad(<5 yrs) prognoses.

Jong Pill Choi et al (2009) [7] they compared the performance of an Artifical Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network combined both ANN and Bayesian Network. The Nine variables of SEER data which were clinically accepted were used as inputs for the networks. The accuracy of ANN (88.8%) and Hybrid Network (87.2%) were very similar and they both outperformed the Bayesian Network. They found the proposed Hybrid model can also be useful to take decisions.

Muhammad Umer Khan et al(2008) [16] they investigated a hybrid scheme based on fuzzy decision trees on SEER data, they performed experiments using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques. They compared the performance of each for cancer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification.

4. Experimental setup

This section consist an experimental work with data mining technique and data mining software. We present a use of decision tree technique on breast cancer data analysis. We can learn various decision rules from this experiment. In this part we will create a model by using decision tree technique. Detail description of decision tree has described in section.

4.1 Breast Cancer Data Source and Description

Breast cancer data was taken from UCI machine learning data repository [28]. This is a secondary data. Dataset consist 10 attributes and 699 instances. Data set description represent in following table.

Table 1 Dataset Description

S.No.	Attribute Name	Range and class value
1	Clump_Thickness	1-10
2	Cell_Size_Uniformity	1-10
3	Cell_Shape_Uniformity	1-10
4	Marginal_Adhesion	1-10
5	Single_Epi_Cell_Size	1-10
6	Bare_Nuclei	1-10
7	Bland_Chromatin	1-10
8	Normal_Nucleoli	1-10
9	Mitoses	1-10
10	Class	Malignant, bengin

4.2Weka

The researcher chose to use these data mining software's mainly due to easy and quick access. Weka provides a number of data mining functionalities such as classification, clustering, association, attribute selection and visualization. Familiarity was also another reason to select Weka data mining software [3]. WEKA has proved itself to be a useful and even essential tool in the analysis of real world data sets. It reduces the level of complexity involved in getting real world data into a variety of machine learning schemes and evaluating the output of those schemes. It has also provided a flexible aid for machine learning research and a tool for introducing people to machine learning in an educational environment. Weka is developed at the University of Waikato in New Zealand. "Weka" stands for the Waikato Environment of Knowledge Analysis [30]. The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and Weka has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. Weka expects the data to be fed into to be in ARFF format.

4.3=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: breast-w
Test mode: evaluate on training data
=== Classifier model (full training set) ===

J48 pruned tree

Cell_Size_Uniformity <= 2
| Bare_Nuclei <= 3.54: benign (406.0/2.0)
| Bare_Nuclei > 3.54
| | Clump_Thickness <= 3: benign (11.0)
| | Clump_Thickness > 3
| | | Bland_Chromatin <= 2
| | | | Marginal_Adhesion <= 3: malignant (2.0)
| | | | Marginal_Adhesion > 3: benign (2.0)
| | | Bland_Chromatin > 2: malignant (8.0)
Cell_Size_Uniformity > 2
| Cell_Shape_Uniformity <= 2
| | Clump_Thickness <= 5: benign (19.0/1.0)
| | Clump_Thickness > 5: malignant (4.0)
| Cell_Shape_Uniformity > 2
| | Cell_Size_Uniformity <= 4
| | | Bare_Nuclei <= 2
| | | | Marginal_Adhesion <= 3: benign (11.0/1.0)
| | | | Marginal_Adhesion > 3: malignant (3.0)
| | | Bare_Nuclei > 2
| | | | Clump_Thickness <= 6
| | | | | Cell_Size_Uniformity <= 3: malignant (13.0/2.0)
| | | | | Cell_Size_Uniformity > 3
| | | | | | Bare_Nuclei <= 6: benign (4.0)
| | | | | | Bare_Nuclei > 6: malignant (7.0/1.0)
| | | | Clump_Thickness > 6: malignant (32.0/1.0)
| | Cell_Size_Uniformity > 4: malignant (177.0/5.0)

Number of Leaves: 14

Size of the tree: 27

=== Summary ===

Correctly Classified Instances 686 98.1402 %
Incorrectly Classified Instances 13 1.8598 %

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.98	0.017	0.991	0.98	0.986	0.989	benign
0.983	0.02	0.963	0.983	0.973	0.989	malignant
Weighted Avg. 0.981	0.018	0.982	0.981	0.981	0.989	

=== Confusion Matrix ===

a	b	<-- classified as
449	9	a = benign
4	237	b = malignant

5. Conclusion and Future Work

In this paper we present the research review in the use of data mining in breast cancer. We observed that neural network and decision approach mostly used by various researchers to create a predictive model and decision rules from the breast cancer data. Most of them they have done the comparative study of algorithm to take breast cancer data. In second part of this paper we conduct an experimental work. We can find various if….then rules from decision tree which is represent in section 4.3.we have used J48 classifier of WEKA which is an extension form of ID3 algorithm of decision tree. In future work we will extend our work. We will try to develop a model for taking critical dieses datasets such as cardio patient and cancer datasets and then compare weka classifier.

References

[1] Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, (2010) "A Data Mining approach for detection of high-risk Breast Cancer groups," *Advances in Soft Computing*, vol. 74, pp. 43-51, 2010.

[2] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed (2010), "Using data mining for assessing diagnosis of breast cnacer," *in Proc. International multiconfrence on computer science and information Technology*, 2010, pp. 11-17.

[3] Bellaachia Abdelghani and Erhan Guven(2006), "Predicting Breast Cancer Survivability using Data Mining Techniques," *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining,*" 2006.

[4] Barlow W. E., White E., Ballard-Barbash R., Vacek, P. M. et al (2006) "Prospective breast cancer risk prediction model for women undergoing screening mammography," J. Natl. Cancer Inst., vol. 98, no. 17, pp. 1204–1214, 2006

[5] Breiman L, Friedman JH, Olshen RA, Stone CJ(1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software; 1984.

[6] Chang Pin Wei and Liou Ming Der,(2010) "Comparision of three Data Mining techniques with Ginetic Algorithm in analysis of Breast Cancer data".[Online]. Available:http://www.ym.edu.tw/~dmliou/Paper/ compar_threedata.pdf

[7] Choi J.P., Han T.H. and Park R.W.(2009), " A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", *J Korean Soc Med Inform*, 2009, pp. 49-57.

[8] Chi C.L., Street W.H. and Wolberg W.H.(2007), "Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets", *Annual Symposium Proceedings / AMIA Symposium*, 2007.

[9] Choi J.P., Han T.H. and Park R.W.(2009), " A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", *J Korean Soc Med Inform*, 2009, pp. 49-57.

[10] Delen Dursun , Walker Glenn and Kadam Amit(2005) , "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine* ,vol. 34, pp. 113-127 , June 2005.

[11] Darius JEGELEVI ˇ CIUS, Ar¯unas LUKOŠEVI ˇ CIUS(2002) , Application of Data Mining Technique for Diagnosis of Posterior Uveal Melanoma ,INFORMATICA, 2002, Vol. 13, No. 4, 455–464

[12] Endo A., Shibata T., and H. Tanaka,(2008) "Comparison of seven algorithms to predict breast cancer survival," Biomedical Soft Computing and Human Sciences, vol. 13 2, pp. 11–16, 2008.

[13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smith. 1996. From Data Mining to Knowledge Discovery: An Overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smith and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pages 1-34. MIT Press, Cambridge, MA.

[14] Gandhi Rajiv K., Karnan Marcus and Kannan S.(2010), "Classification rule construction using particle swarm optimization algorithm for breast cancer datasets," *Signal Acquisition and Processing. ICSAP, International Conference*, 2010, pp. 233 – 237.

[15] Jamarani S. M. h., Behnam H. and Rezairad G. A.(2005), "Multiwavelet Based Neural Network for Breast Cancer Diagnosis", *GVIP 05 Conference*, 2005, pp. 19-21.

[16] Khan M.U., Choi J.P., Shin H. and Kim M (2008), "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", *Conf Proc IEEE Eng Med Biol Soc.*, 2008, pp. 48-51.

[17] Kuo, W.-J., R.-F. Chang, D.-R. Chen and Ch.Ch. Lee (2001). Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast Cancer Research and Treatment*, 66, 51–57.

[18] Lee Heui Chul, Seo Hak Seon and Choi Chul Sang(2001), "Rule discovery using hierarchial classification structure with rough sets," *IFSA World Congress and 20th NAFIPS International Conference*, 2001, vol.1 , pp. 447-452.

[19] Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L.et al(1999) , "Artificial Neural Networks Applied to Survival Prediction in

Breast Cancer", *Oncology International Journal for Cancer Resaerch and Treatment*, vol. 57, 1999.

[20] Padmavati J.(2011), "A Comparative study on Breast Cancer Prediction Using RBF and MLP," *International Journal of Scientific & Engineering Research*, vol. 2, Jan. 2011.

[21] Quinlan J.(1996) Induction of decision trees. Mach Learn 1986;1:81—106.

[22] Quinlan J. (1993) C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.

[23] Rajesh K., Sangeetha V. (2012) Application of Data Mining Methods and Techniques for Diabetes Diagnosis, **International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012 pp** 224 **ISSN: 2277-3754**

[24] Satyanandam N., Satyanarayana Ch. , Md.Riyazuddin,(2012) **Data Mining Machine Learning Approaches and Medical Diagnose Systems : A Survey ,***International Journal of Computer & Organization Trends – Volume2Issue3- 2012 ,PP 53-60* ISSN: 2249-2593 http://www.internationaljournalssrg.org

[25] Sarvestan Soltani A. , Safavi A. A., Parandeh M. N. et al. (2010)., "Predicting Breast Cancer Survivability using data mining techniques," *Software Technology and Engineering (ICSTE), 2nd International Conference*, 2010, vol.2, pp.227-231.

[26] Sudhir D., Ghatol Ashok A., Pande Amol P(2006)., "Neural Network aided Breast Cancer Detection and Diagnosis",*7th WSEAS International Conference on Neural Networks*, 2006.

[27] Testard P.-Vaillant(2010), "The war on cancer," CNRS international magazine, vol. 17, pp. 18–21, 2010.

[28] archive.ics.uci.edu/ml/datasets.html.

[29] Williams, G. and M. Hegland et al (1998). A Data Mining Tutorial. Presented at the Second IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN'98).

[30] WEKA http://www.cs.waikato.ac.nz/ml/weka:

Shiv Shakti Shrivastava received his M.Tech (Computer Science and Engineering) degree in 2009 from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. At present he is research scholar at Mewar University Chittorgarh (Raj). He has published and presented many research papers. Also attended many technical workshops. He has 12 years teaching experience.

Anjali Sant received her M.Sc. & Ph.D. (in 2000) from the Barkatullah University, Bhopal (M.P.). Presently she is working as Head of the department of Basic Science, Bhopal Institute of Technology & Science, Bhopal.(M.P.). She is guiding many research scholars in various universities.

Ramesh Prasad Aharwal received his M.Sc. (Mathematics) degree in 1998 from Dr. H. S. Gour University, Sagar, M.C.A. degree in 2001 from R.G.P.V. Technical university Bhopal and then received Ph.D. degree in 2010 from Barkatullah University Bhopal. He has 12 years teaching experience. Presently he is working as Asstt. Prof. (Mathematics) in Govt. P. G. College, Damoh (M. P.).