Approaches of Classifacation to Policy of Analysis of Medical Data

[1]ANIL RAJPUT, [2]RAMESH PRASAD AHARWAL, [3]NIDHI CHANDEL, [4]DEVENDRA SINGH SOLANKI, [5]RITU SONI

[1]Asstt Prof., Department of Mathematics and Computer Science, Sadhuvaswani College, Bairagarh, Bhopal (M.P.), India
[2]Asstt. Prof., Department of Mathematics and Computer Science, Govt. P.G. College Bareli (M.P.), India

[3]Asstt Prof., Department of Computer Science, Career College, Bhopal (M.P.), India
[4]Asstt. Prof., Department of Mathematics, M.L.C., Govt. Girls P.G. College Khandwa (M.P.), India
[5]Asstt Prof., Department of Computer Science, N.R.I. Group of Instiuts Bhopal (M.P.) India

ABSTRACT**:**
Real life data mining approaches are interesting because they often present a different set of problems for data miners. We have done such real life application in this paper. Application of data mining and knowledge discovery and database techniques are very beneficial but highly challenging in the field of medical and health care. In this study we have used classification techniques for analysis of cancer patient data sets. We have used real datasets in this study. This paper leads the study of data mining in the field of health care.

Keywords: Knowledge Discovery and Database (*KDD), Classification, Data Mining, Decision tree, WEKA, Neural network*

1. INTRODUCTOIN

Data mining is a relatively new field of research whose major objective is to acquire knowledge from large amounts of data. Data mining refers to extracting or mining knowledge from large amounts of data It is a multidisciplinary field bridging many technical areas such as databases technology, statistics, artificial intelligence, machine learning, pattern recognition and data visualization methods In medical and health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available. On the one hand, practitioners are expected to use all this data in their work but, at the same time humans in a short time to make diagnosis, prognosis and treatment schedules, cannot process such a large amount of dataset. Data mining technique has become an established method for improving statistical tools. In this study we use classification techniques for analysis and creating a predictive model. Decision tree, neural network classification techniques are included in this paper. Brief description of data mining is given in next section. But Data Mining is a crucial part of the KDD, and then firstly we will understand about the KDD.

2. KNOWLEDGE DISCOVERY AND DATABASE(KDD)

The research areas of knowledge discovery of databases and data mining have emerged in the recent years with multiple books and various research papers. The Definition of KDD given by [7] as "Knowledge Discovery in Databases in the non trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in Data". In their opinion, a KDD process usually consists of several steps: Data selection, Preprocessing, Transformation, Data mining and interpretation or evaluation of the results. Each step in KDD process has its own meaning and importance. Process of KDD is shows in following figure.

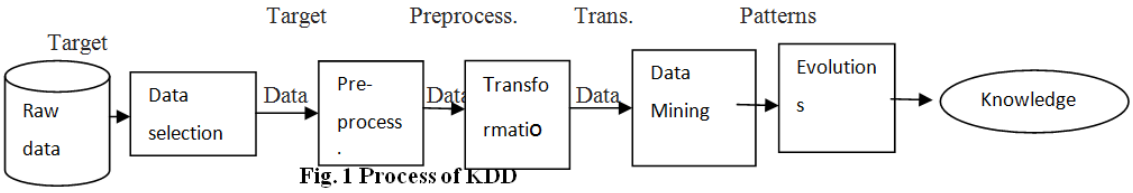


Fig. 1 Process of KDD

Manuscript received November 5, 2009
Manuscript revised November 20, 2009

2.1 DATA MINING

Data Mining is the core task in the process of KDD. It consists of applying computational techniques to extract useful pattern or knowledge from the given datasets. It can also be seen as a combination of tools, techniques and process in KDD. It has six basic functions or activities which classified into two categories. First is direct and second one is indirect. Specifically classification, estimation and prediction are directed where available data is used to build a model. Second category consist association rules, Clustering, description and Visualization which are used for establish some relationship among all variables.

According to author [14], Data mining has two faces: knowledge discovery and decision-making. As a knowledge discovery tool, some data mining algorithms used in this paper which produce explicit knowledge (IF... THEN rules) that can be analyzed by a user. The users may learn new knowledge and at the same time may pose questions to be addressed by a targeted research. The decision-making facet of data mining overlays with decision making and prediction theories, and it produces outcomes of three different types: high confidence decision, low confidence decision, and no-decision.

2.2 CLASSIFICATOIN:

Classification is an important data mining task that analyses a given training set and develop a model for each class according to the features present in the data. There are many approaches used to develop classification model including decision tree, neural network, nearest neighbor method and rough set based method [8].this techniques is mostly used in the field of medical data mining. We may classify patient records with the problem of heart disease. Suppose D is a database of patient which have the set of tuples $\left(t_1, t_2, t_3, \ldots. t_n\right)$ where $t_1, t_2, t_3, \ldots. t_n$ the values of attributes $A_1, A_2, A_3 \ldots\ldots A_n$ of related disease. We may define various classes $C = \left\{C_1, C_2, C_3 \ldots C_m\right\}$ of particular classification types of disease. The classification problem can be defined as a function $F : D \rightarrow$ 잶 where each t_i D is mapped to $f(t_i)$ belonging to some C_j .Example of classification can be seen in [2]. They use the classification task and provide a comprehensive study of classification techniques with more emphasis on classical and incremental decision tree. Example of classification tree is shown on following figure. This figure is adopted from [8].

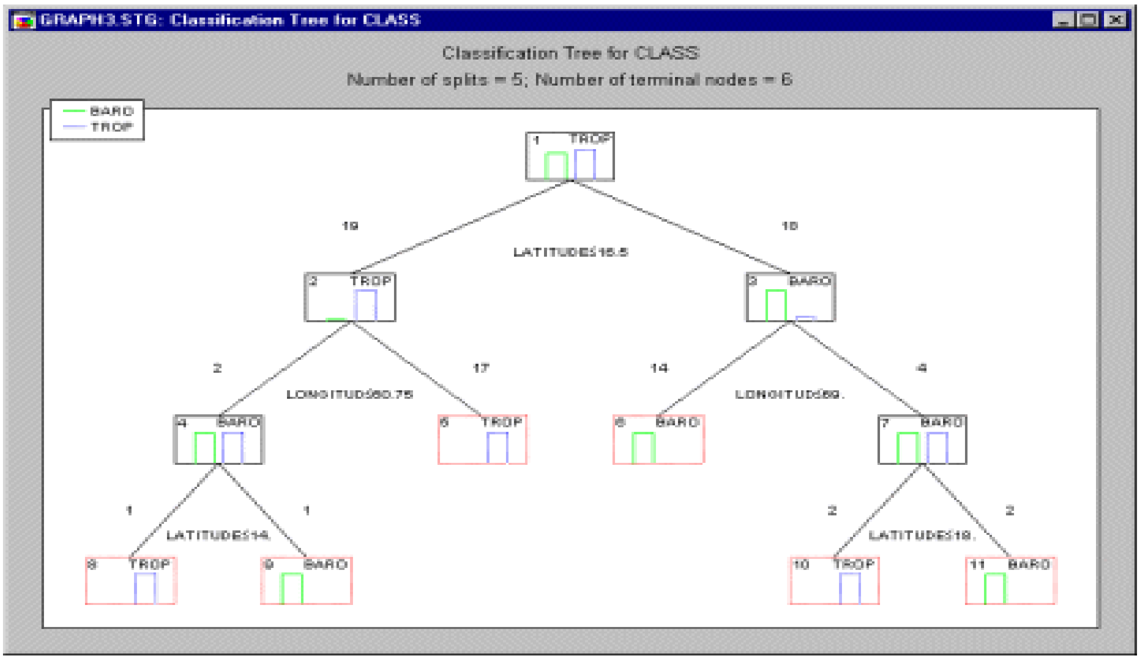


Fig. 2 decision tree

3. DATA SOURCE AND DESCRIPTOIN

Data is taken from medical record office of Gandhi medical college Bhopal (M.P.) India. We have spent at here five days. To access these data from discharge ticket of the cancer patient. discharge ticket consist many fields such as central registration number of patient, patient identification number ,name ,age, sex, blood pressure, pulse rate location, diagnosis, site of cancer date of admission and date of discharge etc, these data is typed in MS Excel. Description of data is summarized in the table 1.

Attribute #	**Instance #**	**Numeric**	**Nominal**	**Class**
5	102	4	2	5

Table1. Description of datasets

3.1 PREPARATION OF THE DATA FROM RAW TO CLEAN DATA:

Having obtained the raw data, it must be massaged into a form suitable for processing by the automated tools. In the case of the WEKA system, the data is extracted and translated into a standard format we call ARFF, for Attribute Relation File Format [9**].** This generally involves taking the physical extract of a database and processing it through a series of steps to generate an ARFF dataset. Weka’s pre-processing capability is encapsulated in an extensive set of routines, called filters that enable data to be processed at the instance and attribute value levels.

4. ISSUES OF SOFTWARE

The **WEKA** (Waikato Environment for Knowledge Analysis) software was developed in the University of New Zealand. A number of data mining methods are implemented in the WEKA software. Some of them are based on decision trees like the J48 decision tree, some are rule-based like ZeroR and decision tables, and some of them are based on probability and regression, like the Naïve Baye’s algorithm. The data that is used for WEKA should be made into the ARFF(Attribute Relation file format) format and the file should have the extension dot ARFF (.arff). WEKA is a collection of machine learning algorithms for solving real world data mining problems. It is written in Java; WEKA runs on almost any platform and is available on the web at www.cs.waikato.ac.nz/ml/weka.

WEKA consist many classifiers which is used for creating a model, pattern and analysis of datasets. We have used J48, multilayer Perceptron evaluate the accuracy of the knowledge generated by data mining algorithms, a 10-fold cross validation was used, where a random 10% of the records were removed and remaining 90% were utilized to generate the rules. The 10% removed were then tested on the generated rule set. This process was repeated 10 times to ensure the generality of the rule sets for future predictions.

5. CLASSIFICATION USING J48 DECISION TREE

This experiment has done in following steps. Details of each step described as follows**.**

5.1 ATTRIBUTE RELEVANCE ANALYSIS

Attribute relevance analysis is used to help identify strong and weak attributes. An attribute is considered strong with respect to a given class if the values of the attribute can be used to distinguish the class from others. The first step in attribute relevance analysis is calculating the information gain.

5.2 CALCULATING INFORMATION GAIN

Let *S* be a set of training samples, where the class label of each sample is known. Each sample is an example of an instance. Let there be *m* classes. One attribute, *A*, is used to determine the class of training samples. Let *S* contain *si* samples of class C_i , for $i = 1, 2, .. \ldots, m.$ an arbitrary sample belongs to class C_i with probability *si /s*, where *s* is the total number of samples in set *S*. The expected information needed to classify a sample is [8]. Information gain is calculated by following formula.

$$I(S_1, S_2, .., S_m) = -\sum_{i=1} \frac{S_i}{S} \log_2 \frac{S_i}{S}$$

An attribute A with values $\{a_1, a_2, \ldots, a_v\}$ can be used to partition S into the subsets $\{S_1, S_2, \ldots, S_v\}$, where S_j contains those samples in S that have value a_j of A. Let S_j contain s_{ij} samples of class C_i . The expected information based on this partition by A is known as the entropy of A. It is the weighted average, shown by [8]

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

Therefore the information gain obtained by this partitioning on defined by [8].

$$Gain(A) = I(S_1, S_2, .., S_m) - E(A)$$

5.3 DECISION TREE ANALYSIS

Decision trees are a useful data analysis tool as they are easy to understand and can be easily transformed into rules. Decision trees are constructed using only those attributes best able to differentiate concepts. The main goal in a decision tree algorithm is to minimize the number of tree levels and tree nodes. The C4.5 [12] decision tree algorithm uses a measure taken from information theory to help with the attribute selection process.

5.4 DECISION TREE GENERATED USING WEKA

The decision tree algorithm that we used in WEKA, J48, gives us an opportunity to control the confidence factor and training sample size (controlled by the cross-validation option). Our objective is to get a decision tree that minimizes the expected error rate, with the highest amount of correctly classified instances. give us the highest amount of correct classification; hence the decision tree was generated with 90% confidence and 10-fold cross validation this decision tree is shown in Figure 5. In WEKA, the confidence factor is used to address the issue of tree pruning. When a decision tree is being built, many of the branches will reflect anomalies due to noise or outliers in the training data. Tree pruning uses statistical Measures to remove these noise and outlier branches, allowing for the confidence factor. This means that our dataset did not have much noise or outlier cases, so there was not much to prune faster classification and improvement in the ability of the tree to correctly classify independent test data (8 Han & Kamber, 2006). A smaller confidence factor will incur more pruning, so for example if a 98% confidence factor is used, our tree will incur less pruning. We ran WEKA with a very wide range of confidence factors, but the results were not reacting to.

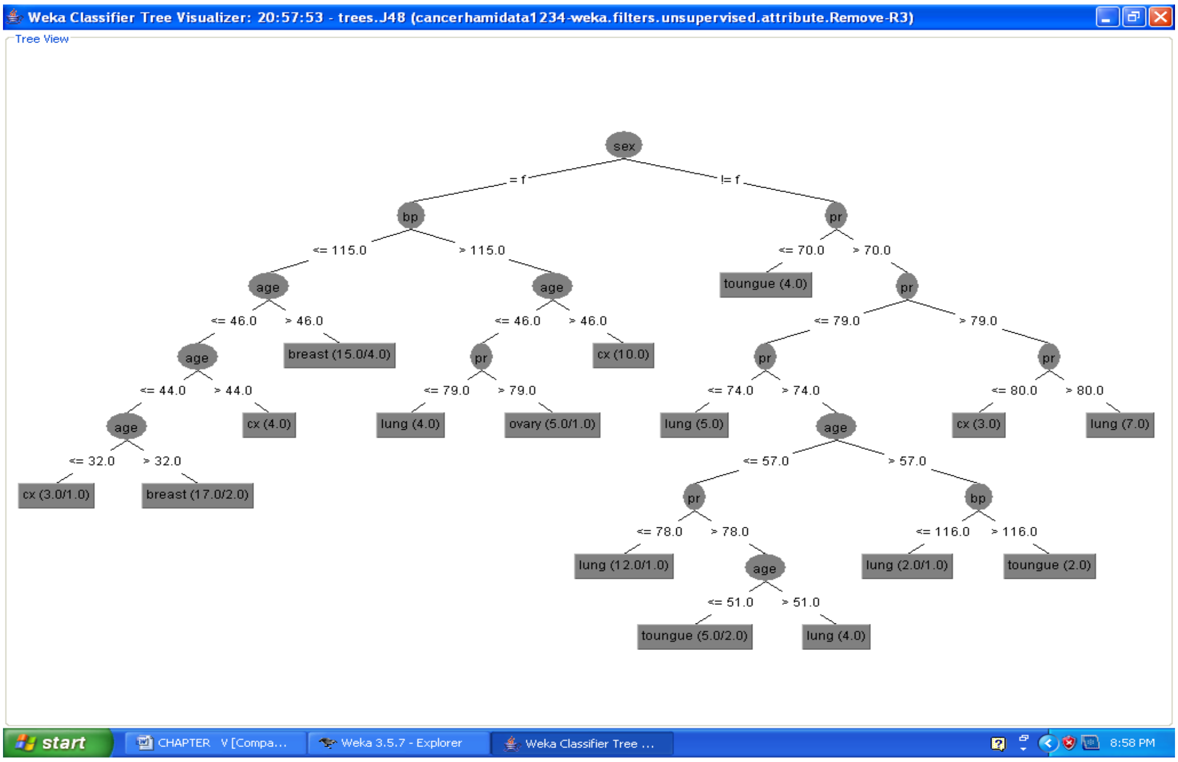
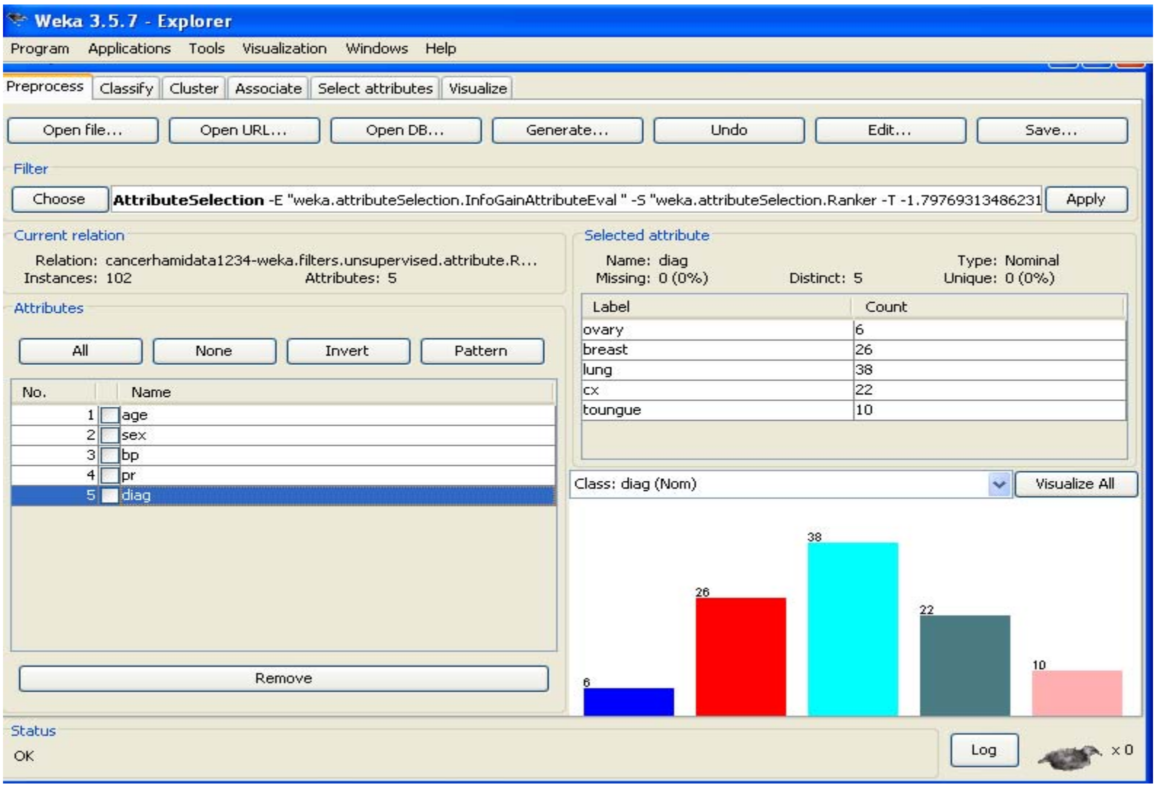


Fig. 3 Decisoin tree which is created with WEKA

Some Screen shots which is generated during this experiment.


Weka 3.5.7 - Explorer
Program Applications Tools Visualization Windows Help
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose
Apply
Current relation
Attributes
All None Invert Pattern
Remove
Selected attribute
Visualize All
Status
Log
x 0
6
26
38
22
10

AttributeSelection -E "weka.attributeSelection.InfoGainAttributeEval " -S "weka.attributeSelection.Ranker -T -1.79769313486231

Relation: cancerhamidata1234-weka.filters.unsupervised.attribute.R...
Instances: 102 Attributes: 5

No.	Name
1	□ age
2	□ sex
3	□ bp
4	□ pr
5	□ diag

Name: diag Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

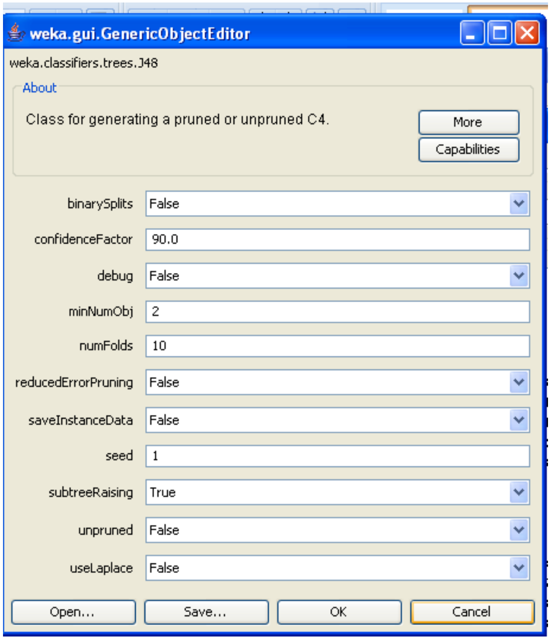
Label	Count
ovary	6
breast	26
lung	38
cx	22
toungue	10

Class: diag (Nom)

OK

Fig. 4 Class Distribution of each Class

Fig. Class Distribution of each attribute.

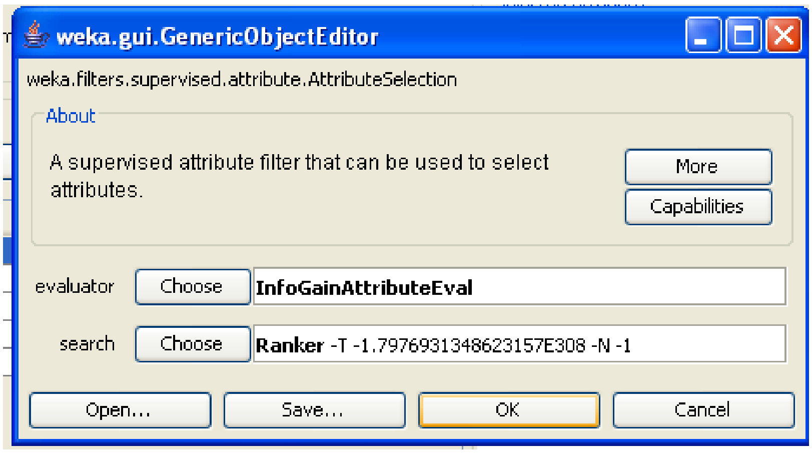


weka.classifiers.trees.J48

Class for generating a pruned or unpruned C4.

binarySplits	False
confidenceFactor	90.0
debug	False
minNumObj	2
numFolds	10
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False

Fig 5. Parameters which is taken for createing decisoin tree in J48



weka.gui.GenericObjectEditor

weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes.

evaluator: InfoGainAttributeEval

search: Ranker -T -1.7976931348623157E308 -N -1

Fig 6. Parameters which is taken for Preprocessing

5.5 WEKA GENERATED OUTPUT FOR J48 ALGORITHM

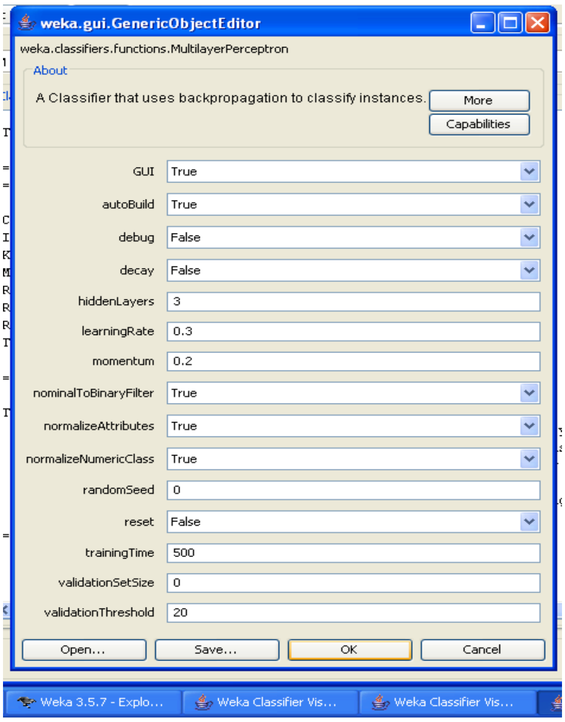
```
=== Run information ===

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 unpruned tree
------------------

sex = f
|  bp <= 115.0
|  |  age <= 46.0
|  |  |  age <= 44.0
|  |  |  |  age <= 32.0: cx (3.0/1.0)
|  |  |  |  age > 32.0: breast (17.0/2.0)
|  |  |  age > 44.0: cx (4.0)
|  |  age > 46.0: breast (15.0/4.0)
|  bp > 115.0
|  |  age <= 46.0
|  |  |  pr <= 79.0: lung (4.0)
|  |  |  pr > 79.0: ovary (5.0/1.0)
|  |  age > 46.0: cx (10.0)
sex != f
|  pr <= 70.0: toungue (4.0)
|  pr > 70.0
|  |  pr <= 79.0
|  |  |  pr <= 74.0: lung (5.0)
|  |  |  pr > 74.0
|  |  |  |  age <= 57.0
|  |  |  |  |  pr <= 78.0: lung (12.0/1.0)
|  |  |  |  |  pr > 78.0
|  |  |  |  |  |  age <= 51.0: toungue (5.0/2.0)
|  |  |  |  |  |  age > 51.0: lung (4.0)
|  |  |  |  age > 57.0
|  |  |  |  |  bp <= 116.0: lung (2.0/1.0)
|  |  |  |  |  bp > 116.0: toungue (2.0)
|  |  pr > 79.0
|  |  |  pr <= 80.0: cx (3.0)
|  |  |  pr > 80.0: lung (7.0)

Number of Leaves  :       16

Size of the tree :   31
```

6. CLASSIFICATION WITH NEURAL NETWORK

According to Author [16] Data mining tasks can be classified into two categories: Descriptive and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea. Predictive data mining allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered form the database. Artificial Neural Network is one of many data mining analytical tools that can be utilized to make predictions on key healthcare indicator such as cost or facility utilization. Artificial Neural networks are well suited to tackle problems that people are good at solving, like prediction and pattern recognition. Neural networks are known to produce highly accurate results and in medical applications, can lead to appropriate decisions. It has been applied within the medical domain for clinical diagnosis, image analysis and interpretation [8]. In this study we realized the classification model with back propagation, which is the most popular neural network learning algorithm.

6.1 BACK-PROPAGATION NEURAL NETWORKS

A back-propagation network trains by using a random initialization of weights describing a set of partitions; an error surface is iteratively minimized by successively considering the error relative to each input point many times. This is a gradient-descent method, so therefore back-propagation networks are prone to issues with local minima in the error space. The main obstacle to their application in this domain is the resulting lack of interpretability of the final stable state. Each node in the hidden layer (or layers) of a back-propagation network allows a greater degree of non-linearity in the final partition space by controlling the location and angle of some high-dimensional hyper-plane. While the geometry of these planes is accessible through an examination of the weights, the actual topology of the space is not easily visualizable, and certainly is not explainable to a user not familiar with the mathematics involved.



weka.gui.GenericObjectEditor

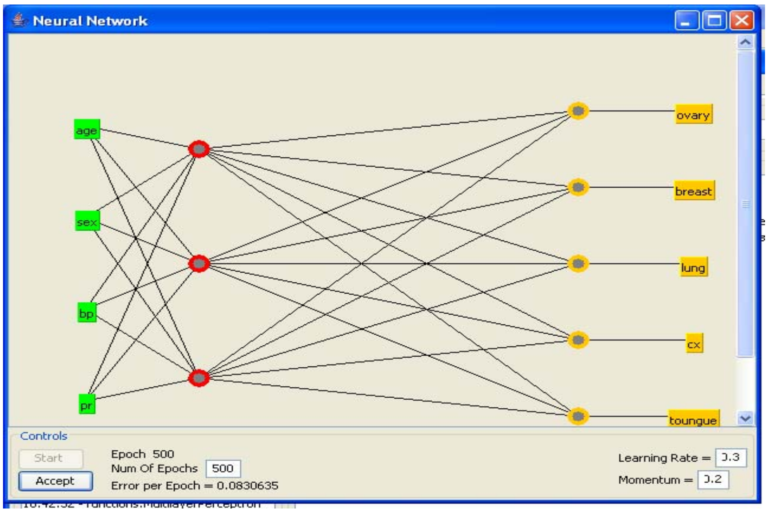
weka.classifiers.functions.MultilayerPerceptron

About

A Classifier that uses backpropagation to classify instances.

GUI	True
autoBuild	True
debug	False
decay	False
hiddenLayers	3
learningRate	0.3
momentum	0.2
nominalToBinaryFilter	True
normalizeAttributes	True
normalizeNumericClass	True
randomSeed	0
reset	False
trainingTime	500
validationSetSize	0
validationThreshold	20

Fig. 7 Paramets for Backpropagation neural network



Neural Network

Epoch 500
Num Of Epochs 500
Error per Epoch = 0.0830635

Learning Rate = 0.3
Momentum = 0.2

Fig. 8 Structure of Neural network Generated with WEKA

WEKA GENERATED OUTPUT FOR MULTILAYE PERCEPTRON ALGORITHM

```
Scheme:
weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M
0.2 -N 500 -V 0 -S 0 -E 20 -H 3 -G -R
Relation:     cancerhamidata1234-
weka.filters.unsupervised.attribute.Remove-R3
Instances:    102
Attributes:   5
              age
              sex
              bp
              pr
              diag

Test mode:    split 66% train, remainder test

=== Classifier model (full training set) ===

Sigmoid Node 0
    Inputs    Weights
    Threshold    -2.0308105126087814
    Node 5    -0.4218634220133945
    Node 6    0.17909785296623998
    Node 7    -2.368127885893712
Sigmoid Node 1
    Inputs    Weights
    Threshold    -7.4964051458082634
    Node 5    4.799153828562641
    Node 6    2.5201271363927003
    Node 7    5.145188537075538
Sigmoid Node 2
    Inputs    Weights
    Threshold    -0.17678085811395544
    Node 5    -5.369097374531843
    Node 6    -3.507744539571276
    Node 7    3.2456088365502613
Sigmoid Node 3
    Inputs    Weights
    Threshold    -3.7723541902515443
    Node 5    -3.2881753391780886
    Node 6    7.260810497751853
    Node 7    -6.125208219216489
Sigmoid Node 4
    Inputs    Weights
    Threshold    -1.2839250398486435
    Node 5    4.786148180565925
    Node 6    -6.865131901801936
    Node 7    -5.569809073576164
Sigmoid Node 5
    Inputs    Weights
    Threshold    -4.149915098003342
    Attrib age    -5.853874459399585
    Attrib sex    -0.9460857235117729
    Attrib bp    -3.9230027198388524
    Attrib pr    -11.499473003959762
Sigmoid Node 6
    Inputs    Weights
    Threshold    -5.3152267134730735
    Attrib age    7.7177313555827585
    Attrib sex    -7.2846781315437505
    Attrib bp    -4.471206880947062
    Attrib pr    -3.7265150496261437
Sigmoid Node 7
    Inputs    Weights
    Threshold    -7.40169879633609
    Attrib age    12.478100781181716
    Attrib sex    1.4198124391620117
    Attrib bp    -11.660606255186122
    Attrib pr    -1.1426911434116636
Class ovary
    Input
    Node 0
Class breast
    Input
    Node 1
Class lung
    Input
    Node 2
Class cx
    Input
    Node 3
Class toungue
    Input
    Node 4


Time taken to build model: 3.88 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances              91.5714 %
Incorrectly Classified Instances             08.4286 %
=== Detailed Accuracy By Class ===
```

7. CONCLUSION:

In this paper we have described classification techniques for cancer datasets. We have used data mining classifiers to generate decision tree and neural network. In this paper we have used WEKA software for our experiment. Experimental result is summarized fig. 3. This paper leads the study of data mining in health care datasets.

REFERENCES:

[1] R. Agrawal, T. Imielinski, et al, “Database Mining: A Performance Perspective”, IEEE Transactions on Knowledge and Data Engineering, pp. 914-925. 1993.

[2] Sultan Ahmed, AI. Hegam, “Classical and incremental classification in data mining process “ International Journal of Computer science and Network security, Vol. 7, no12. , 2007.

[3] D.S., Barr, G.Mani ,”Using Neural Nets to Manage Investments”, AI Expert, February, pp. 16-21. 1994.

[4] L.Breiman, J.H. Friedman, R.Olshen, et al. Classification and Regression Tree Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California 1984.

[5] S.Chuddari, “Data Mining and Database Systems : Where is the Intersection?”, IEEE Bulletin of the Technical Committee on Data Engineering, Vol.21 No.1, pp. 4-8. 1998.

[6] M. Chen, J., Han, P.S. Yu “Data Mining: An Overview from Database Perspective”, IEEE Transactions on Knowledge and Data Engineering, Vol. 8 No.6. 1996,

[7] U.M.Fayyad., G, Piatetsky, P. Shapiro, et.al., “Advances in knowledge Discovery and Data Mining “ AAAI/MIT Press.,1996.

[8] J.Han and M .Kamber, “Data Mining: concept and techniques “first edition, Harcoart India private Limited. 2001.

[9] G. Holmes, A. Donkin, and I.H., Witten, “Weka: a machine learning workbench.” Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, pp. 357- 361, 1994.

[10] M. Kantaradzic, Data Mining: Concepts, Models, Methods, and Algorithms, IEEE Press and J ohn Wiley, New York, NY. 2003.

[11] Harleen Kaur and Siri Krishan Wasan, “Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science 2 (2): 194-200, ISSN 1549-3636, 2006.

[12] J. R Quinlan,. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos

[13] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning representations by back-propagating errors. Nature, 323, 533–536, 1986.

[14] C.Shital Shaha, et.al., “Patient-recognition data-mining model for BCG plus interferonimmunotherapy bladder cancer treatment” Elsevier, Computers in Biology and Medicine vol. 36 pp 634–655. 2006.

[15] M .Stone,Cross-validatory choice and assessment of statistical predictions, J. Royal Stat. Soc. 36, pp 111–147. 1974

[16] Serhat Ozekes and Onur Osman, “Classification and Prediction in Data Mining with Neural Networks”, “Journal of Electrical and Electronics Engineering “,Vol. 1. No. 3, pp 707-712. , 2003.

[17] T.B Teshma et al “Data Mining using adaptive regression tree” international Journal of Simulation vol. 6 no.10 and 11 ISSN 1473-804 online. 2007