INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

DISTRICT OF NILGIRIS RAINFALL LEVEL ANALYSIS USING DIFFERENT CLASSIFICATION TECHNIQUES IN WEKA

P.Rakkimuthu [1], G.D.Praveenkumar [2], M.Subha [3].

[1]Research scholar, kaamadhenu arts and Science College, Erode, E-Mail: rcbrakky@gmail.com,
[2]Assistant.Professor, Bharathiar university arts and Science College, Gudalur, E-Mail: erodegd@gmail.com,
[3]Assistant.Professor, kaamadhenu arts and Science College, E-Mail: kasc.subha@gmail.com.

Abstract: -The goal of data mining is to extract or knowledge from large amount of data. This paper deals with the analysis of rainfall level in Nilgiri district using classification technique in weka. The weka tool is open source software. The rainfall data sets are collected from the official website. The data set are converted into proper preprocessing format then implementing different classification algorithms in weka, to calculate processing time, evaluate training set, average weight, confusion matrix.

Keywords: Data mining, data set, classification, confusion matrix.

1. INTRODUCTION

The mountain area of nilgiri district found in Western Ghats. The mountain spread across the division between the states of TamilNadu, kerela Karnataka. The general climatic condition in the district is cool. In TamilNadu nilgiri is highest rainfall area and normal annual rainfall over the district varies over a wide range from 560.mm to 886.6mm. The chances of receiving normal annual rainfall vary from 40% to 50% over the district. The rainfall data set are used to calculate the rainfall level in during the year of 2000 to 2014

1.1 Data Mining

Data mining is the process of automated data Analysis techniques to discover previously undetected relationship among data items. Data mining often involves the analysis of data stored in data ware house.

There are many data mining techniques are available like

- Classification
- Clustering
- Pattern recognition
- Association

The data mining gather the data, while the machine learning algorithm are used for taking decision based on the data collected .

1.2 Classification

Classification is data mining technique used to predict group member ship data instances for example for classification to predict whether on a particular day will be sunny, rainy, and popular. Classification technique includes decision tree and neural network. All approaches to perform classification assume some knowledge of data. Often a training set is used to develop the specific parameters required by the technique. Training data consist of sample input data as well as the classification assignment for the data. Domain experts may also be used assist in the process.

2. PROBLEM STATEMENT

The problem is here to discuss about different classification algorithms used to Analysis of rainfall dataset.

3. METHODOLOGY

3.1Data mining process

The following processes of data mining are,

- Selection
- Preprocessing
- Transformation
- Data mining
- Evaluation

3.1.1 Overview of weka tools

The data pre-processing and data mining was performed using the world famous Weka Data Mining tool. Weka is a collection of machine learning algorithms for data mining tasks. Weka is open source software for data mining under the GNU General Public License. This system is developed at the University of Waikato in New Zealand. "Weka" stands for the Waikato Environment for Knowledge Analysis. Weka is freely available. Waikato.ac.nz/ml. weka. The system is written using object oriented language java. Weka provides implementation of state-of-the-art data mining and machine learning algorithm. User can perform association, filtering, classification, clustering, visualization, regression etc., by using weka tool.

3.2 Data Preparation

The data set of rainfall level in nilgiri district research work was downloaded from the website www.idm.gov. The data set one creating an two dimensional spread sheet or database table for weka the data set should have in the format of csv or .arff file format.

3.3 Data Set Preprocessing

The data set are basic concept of data mining. A data set contained less number of attributes and instance only the during the year of 2000 -2014.In this Rainfall dataset divided into 4 class, class A is heavy rainfall, class B is average rainfall, class C below average rainfall and class D is low rainfall.

3.4 Performing Classification on Weka

For performing classification in weka tool, the data set are loaded open file tab in explorer window then performing filter process using preprocessing method to reduce the error , that data set are ready to applied in classify tab .The classification process involves following steps,

- Creating training dataset
- Identify class attribute and classes
- Identify useful attributes for classification

- Learn a model using training examples in training set
- Use the model to classify the unknown data samples

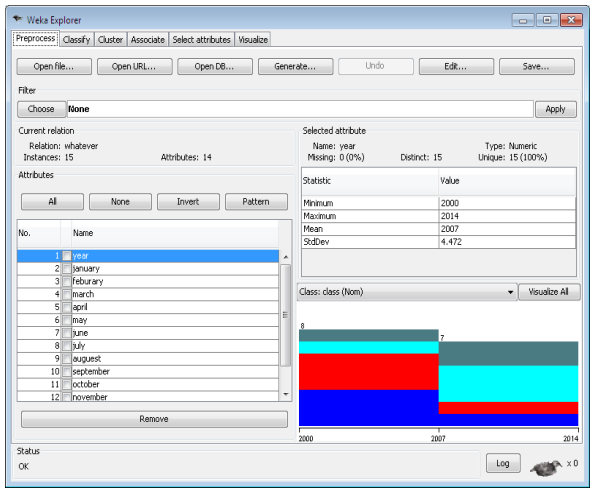

Figure1: Loading data into the file

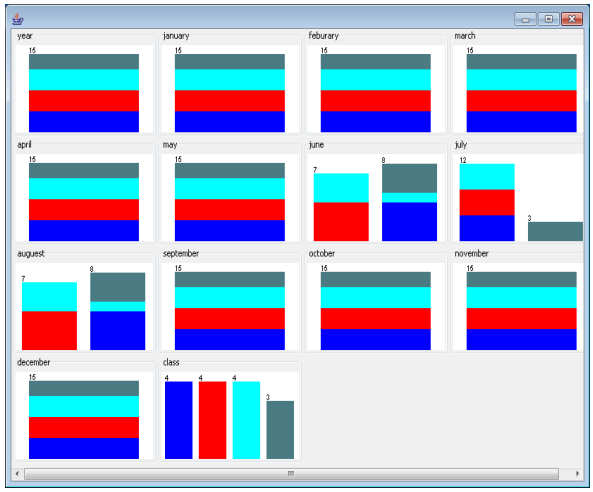

Figure 2: visualize the overall rainfall level in Nilgiri district.

3.4.1 Naive Bayes Classification

Navie Bayes is a simple technique for constructing classifier. Navie Bayes model assigns class label to problem instance represented as vector features values, where the class labels are finite set.

3.4.1.1 Weka processing in Naive Bayes Classification

Scheme: weka.classifiers.bayes.NaiveBayes
Relation: whatever-weka.filters.supervised.attribute.Discretize-Rfirst-last
Instances: 15
Attributes: 14
Test mode: 5-fold cross-validation
Classifier model:
Naive Bayes Classifier:

	Class			
Attribute	c	a	b	d

(0.26) (0.26) (0.26) (0.21)
Time taken to build model: 0.2 seconds
Evaluation on training set :
Correctly Classified Instances 11 73.3333 %
Incorrectly Classified Instances 4 26.6667 %
Detailed Accuracy By Class:

TPRATE	FPRATE	Precision	Recall	F-measure	ROC AREA	CLASS
1	0.182	0.667	1	0.8	1	C
1	0.182	0.667	1	0.8	0.83	A
0	0	0	0	0	0.648	B
1	0	1	1	1	1	D
Weighted Avg.	0.733	0.097	0.556	0.733	0.627	0.861

Confusion Matrix:

```
 a b c d   <-- classified as
 4 0 0 0 | a = c
 0 4 0 0 | b = a
 2 2 0 0 | c = b
 0 0 0 3 | d = d
```

3.4.2 J48 TREE

It builds the decision tree from labeled training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. To make the decision the attribute with highest normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class.

3.4.2.1 Weka processing J48 TREE

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: whatever-weka.filters.supervised.attribute.Discretize-Rfirst-last
Instances: 15
Attributes: 14
Test mode: 5-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
---
june = '(-inf-183.9]': a (7.0/3.0)
june = '(183.9-inf)'
| july = '(-inf-353]': c (5.0/1.0)
| july = '(353-inf)': d (3.0)
Number of Leaves: 3
Size of the tree: 5
Time taken to build model: 0.1 seconds

Evaluation on training set :
Correctly Classified Instances 11 73.3333 %
Incorrectly Classified Instances 4 26.6667 %

Kappa statistic 0.6429

Detailed Accuracy By Class:

TPRATE	FPRATE	Precision	Recall	F-measure	ROC AREA	CLASS
1	0.182	0.667	1	0.8	0.841	C
1	0.182	0.667	1	0.8	0.83	A
0	0	0	0	0	0.42	B
1	0	1	1	1	1	D
Weighted Avg.	0.733	0.097	0.556	0.733	0.627	0.758

Confusion Matrix:

a b c d <-- classified as
4 0 0 0 | a = c
0 4 0 0 | b = a
2 2 0 0 | c = b
0 0 0 3 | d = d

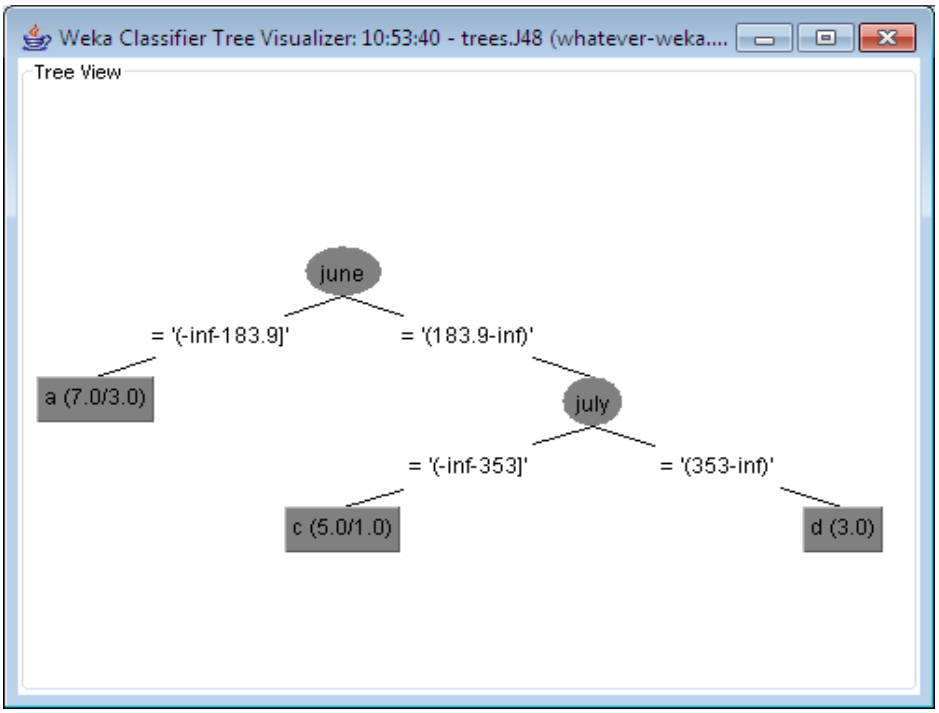

Figure 3: weka.classifiers.trees.J48

3.4.3 LADTREE

Logical Analysis of data is the method for classification proposed in optimization literature. It builds a classifier for binary target variable based on learning a logical expression that can distinguish between positive and negative samples in a data set. The basic assumption of LAD model is that a binary point covered by some positive patterns, but not covered by any negative pattern is positive, and similarly, a binary point covered by some negative patterns, but not covered by positive pattern is negative. The construction of Lad model for a given data set typically involves the generation of large set patterns and the selection of a subset of them that satisfies the above assumption such that each that pattern in the model satisfies certain requirements in terms of prevalence and homogeneity

3.4.3.1 Weka processing LADTREE

Scheme: weka.classifiers.trees.LADTree -B 10
Relation: whatever-weka.filters.supervised.attribute.Discretize-Rfirst-last
Instances: 15
Attributes: 14
Test mode: evaluate on training data
=== Classifier model (full training set) ===
weka.classifiers.trees.LADTree:

: 0,0,0,0
| (1)july = '(-inf-353]': 0.333,0.333,0.333,-1
| | (2)june = '(-inf-183.9]': -0.949,1.068,0.564,-0.683
| | | (4)august = '(-inf-209.65]': -0.491,0.641,0.301,-0.451
| | | (4)august = '(209.65-inf)': -0.542,-1.439,2.482,-0.501
| | (2)june = '(183.9-inf)': 1.875,-0.949,-0.243,-0.683
| | | (3)august = '(-inf-209.65]': -2.998,-0.781,4.531,-0.753
| | | (3)august = '(209.65-inf)': 2.401,-0.793,-0.84,-0.768
| (1)july = '(353-inf)': -1,-1,-1,3
| (5)august = '(-inf-209.65]': -0.813,0.704,0.905,-0.796
| (5)august = '(209.65-inf)': 0.004,-1.376,0.245,1.127
| | (7)july = '(-inf-353]': 0.74,-0.562,0.376,-0.553
| | | (8)june = '(-inf-183.9]': -0.379,-0.377,1.132,-0.376
| | | (8)june = '(183.9-inf)': 1.127,-0.375,-0.377,-0.376
| | (7)july = '(353-inf)': -0.383,-0.376,-0.377,1.136
| (6)june = '(-inf-183.9]': -0.4,0.424,0.371,-0.394
| (6)june = '(183.9-inf)': 0.337,-0.611,-0.27,0.544
Legend: c, a, b, d
#Tree size (total): 25
#Tree size (number of predictor nodes): 17
#leaves (number of predictor nodes): 11
#expanded nodes: 94
#processed examples: 617
#Ratio e/n: 6.5638297872340425

Time taken to build model: 0.02 seconds

Evaluation on training set:
Correctly Classified Instances 13 86.6667 %
Incorrectly Classified Instances 2 13.3333 %

Detailed Accuracy By Class:

TPRATE	FPRATE	Precision	Recall	F measure	ROC AREA	CLASS
1	0	1	1	1	1	C
1	0.182	0.667	1	0.8	0.909	A
0	0	1	0.5	0.667	0.909	B
1	0	1	1	1	1	D
Weighted Avg.	0.867	0.048	0.911	0.867	0.858	0.952

Confusion Matrix:

a b c d <-- classified as
4 0 0 0 | a = c
0 4 0 0 | b = a
0 2 2 0 | c = b
0 0 0 3 | d = d

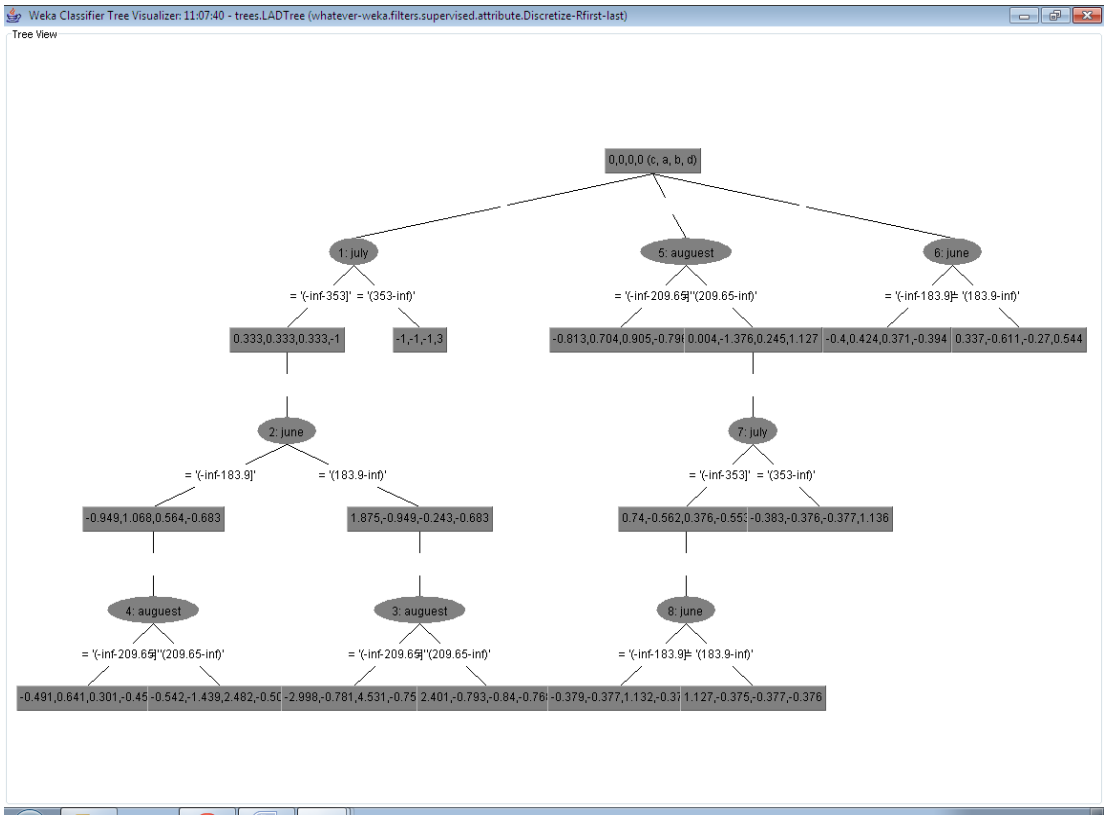

Figure 4:weka.classifiers.trees.LADTree

4. CONCLUSION

The rainwater is the primary source of every living thing. The nilgiri district is highest rainfall area in TamilNadu .Due to climatic change, global warming, afforestation are the problem of decreasing rainfall level year by year in Nilgiri District. The government takes many actions for reduce that factor to improve the rainfall level. Rainfall data set are processed in different classification algorithms developed by weka tool. The data set attributes are converted in proper format of ".arff ".Each classification algorithms present and achieve a high rate of accuracy. It classify the data into the correctly and incorrectly instance.

REFERENCES

[1] An Implementations of ID3: Decision Tree Learning Algorithm Wei Peng,Juhua Chen and Haiping Zhou Project of Comp 9417 : Machine Learning University of New South Wales,School of Computer Science & Engineering,Sydney,NSW 2032,and Australia.

[2] Han,J.,Kamber,M.,Jian P., Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann publishers, 2011.

[3] Hall,M., Frank,E., Holmes,G.,Pfahringer,B.,Reutemann P.,Witten,J.,H., The WEKA data mining software: an update,ACM SIGKDD Explorations Newsletter,V.11 n.1,June 2009 [doi>10.1145/1656274.1656278].

[4] Hornik,K.., Buchta,C., Zeileis ,A., Open-Source Machine learning: RMeets Weka, Journal of Computational Statistics_Proceedings of DSC 2007, Volume 24 Issue 2,May 2009[doi>10.1007/s00180-Conference on Automatic Control, Modelling & Simulation,2010.
[5] www.junauza.com/2010/11/free-data-mining software.html
[6] King,M.,A.,and Elder,J.,F., Evaluation of Fourteen Desktop Data Mining Tools, in Proceedings of the 1998 IEEEE International Conference on Systems, Man and Cybernetics,1998.
[7] Giraud-Carrier, C., and Povel,).,Characterizing Data mining software, intelligent data analysis,v.7n.3,p.181-192,august 2003
[8] weka , the university of Waikato, available at http://www.cs.wailkato.ac.nz/ml/weka/,(accessed 20 April 2011)
[9] uci machine learning repository, available at http://http://archive.ics.uci.edu/ml/,(accessed 22 April 2011)

BIOGRAPHIES

1st **P.RakkiMuthu** is currently working in Assistant Professor in Department of Computer science in Bharathiyar University Arts and Science College, Gudalur. He has completed M.Sc. (CT) in Kongu Engineering, Erode. Currently he is doing his research (Part-time) at Kaamadhenu Arts and Science College, Erode. His areas of interest are Data Mining, Networking.

2nd **G.D.Praveenkumar** is currently working in Assistant Professor in Department of Computer science in Bharathiyar University Arts and Science College, Gudular. He has completed MSC (IT) in Kongu Engineering, Erode. He has published 2 papers in National Conference and published 1 paper in international journals. His areas of interest are Data Mining, Mobile Computing.

3rd **M. Subha** is currently working as an assistant professor in kaamadhenu Arts and Science College, Sathyamangalam. She received her M.Phil. (CS) degree in Bharathiar University, papers in various international journals. Her area of interest is in Data Mining & warehousing, Biometrics, data Structure, Digital image processing.