International Journal of Computer Network and Security(IJCNS)
Vol5. No.1 – Jan-March 2013 Pp. 12-15
©gopalax Journals, Singapore
available at : www.ijcns.com

COLLATIVE STUDY OF CLASSIFIERS IN PATTEREN RECOGNITION

Namratha M[1], Prajwala T R[2],Malvika M[3]
[1] namratha.july@gmail.com, [2] prajwalatr@gmail.com, [3] malusunrise@gmail.com
Mtech in Software Engineering, Dept. of ISE,
PESIT (Peoples Education Society Institute of Technology), Bangalore, India

ABSTRACT

Pattern recognition which is a field of machine learning has evolved from artificial intelligence. In turn it helps us to make decisions and recognize various patterns. The two fields of pattern recognition are classification and regression. Classification a form of supervised learning helps us to classify the training data into correctly labeled dataset. Classification trees also known as classifiers are used to classify data into expected class. Random forest and REPtree are two such classifiers .Random forest is ensemble of predictor variables which classifies the data in input vector into correctly identified classes. REPTree is fast learning regression tree which is suitable for classifying numerical values. This paper presents algorithm and flow chart for classification using random forest and REPTree. Each of these classifiers have their own advantages and disadvantages Experiments are conducted using WEKA3.7 open source tool and IRIS database set. Based on the conclusions drawn from experiment we can choose a suitable classifier.

Keywords : Decision trees, machine learning, pattern recognition, classifiers, random forest , REPTree

1. INTRODUCTION

Machine learning has become one of main stays of information technology. Machine learning has evolved from broad field called artificial intelligence which aims to mimic intelligent abilities of humans by machines. It provides answer to question of how to make machines able to learn. Machine learning is defined as development of computer programs that can teach themselves to grow and change when exposed to new data.

Pattern recognition[1] a field of machine learning is a way of recognizing a pattern using a machine like a computer. It is a study of how machines can observe the environment, learn distinguished pattern from their background and make reasonable decisions about categories of pattern. Thus the two main fields of pattern recognition are classification and regression. . Pattern recognition is a statistical approach of that is used for supervised or unsupervised classification.

Classification[1] is a technique in pattern recognition which deals with supervised learning. One of example of classification technique is classifying email as spam and not spam. In general classification is a problem of identifying to which set of sub-population a new observation belongs to,based on training dataset. A supervised learning system that performs classification is often called a learner or a classifier. A classifier is feed with training data in which each item is correctly labeled. This data is used to train learning algorithm, which creates models that classify similar data.

There are many classifiers like navie bayes classifier,support vector machines,perceptions etc. Classification tree is also one of the classifier. It is used to predict membership of objects in classes of dependent variable based on predictor variable. It is a powerful way to identify multilevel interactions. It is composed of branches which represent attributes and leaves which represent decisions.

2. CLASSIFICATION TREE

Classification tree[1] also called as decision tree maps the observations about an item to conclusions about the target item value wherein we

International Journal of Computer Network and Security(IJCNS)
Vol5. No.1 – Jan-March 2013 Pp. 12-15
©gopalax Journals, Singapore
available at : www.ijcns.com

predict the value of target variable based on input values. The goal is to ensure that the data belongs to expected class. An example is classifying the test data set of the Iris database into the following categories like Iris-Setosa , Iris-Versicolor , iris-Virginica.

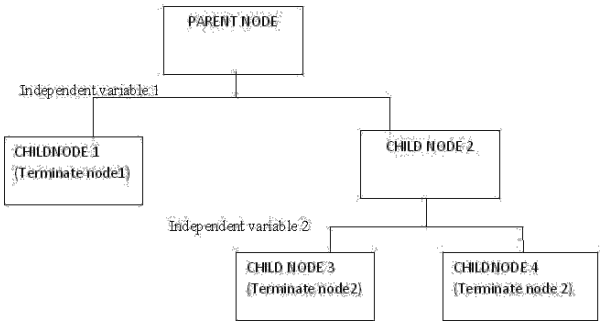

Figure.1. Diagramatic representation of classification tree

In the above diagram the internal nodes (childnode1, childnode2) represent attributes, the leaf nodes(childnode3,childnode4) represent the decisions. Properties of classification tree:

- The hierarchical nature of classification tree helps organize the dataset. The final decision depends on previous dataset.
- The flexibility of classification tree ensures that effects of predictor variables are taken one at a time rather than all at once.
- Classification trees are used to select the attribute which is useful for classifying examples based on entropy or information gain.

3. RANDOM FOREST:

It is a ensemble of tree predictors such that each tree depends on the value of a random vector sampled independently and with same distribution for all trees in fores[2]t. We classify a new object from an input vector by putting the input vector down each of the trees in the forest. Each tree gives a classification and tree votes for that class. The forest with maximum votes is chosen for classification.

Learning algorithm[3]: the iterative procedure is as follows:

1. Let N be number of training cases and M be number of predictor variables.
2. Let m be number of input variables used to determine the decision at a node of tree such that m<M.
3. Choose n of N training set as bootstrap sample to determine accuracy. Use the remaining cases to estimate the error of the tree.
4. For each node randomly assign m predictor variables and compute the best split.
5. Each tree is not pruned and fully grown.

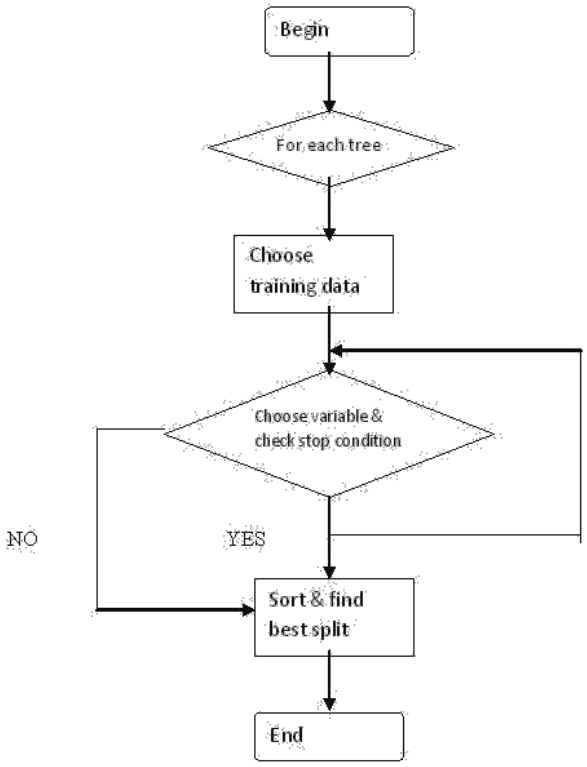

Figure 2:Flow chart for learning algorithm

Features[3]:

- It estimates which variable is important for classification.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers.
- It identifies variable interactions.
- Effective method for estimating missing data and maintains accuracy even when large proportion of data is missing.

Advantages of random forest:

- Runs efficiently for large databases.

International Journal of Computer Network and Security(IJCNS)
Vol5. No.1 – Jan-March 2013 Pp. 12-15
©gopalax Journals, Singapore
available at : www.ijcns.com

- Can handle thousands of input variables without variable deletion
- Results can be saved for future use.

Disadvantages of random forest:

- This technique is not suitable for classifying noisy dataset.
- Random forest is favorable for attributes with more levels for categorical data variables.

Experimental results for IRIS dataset in WEKA:

The training dataset was run using random forest classifier to obtain following result:

```
=== Summary ===

Correctly Classified Instances         150               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0102
Root mean squared error                  0.0554
Relative absolute error                  2.3    %
Root relative squared error             11.7473 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      35.7778 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1        0        1          1       1          1        1         1         Iris-setosa
               1        0        1          1       1          1        1         1         Iris-versicolor
               1        0        1          1       1          1        1         1         Iris-virginica
Weighted Avg.  1        0        1          1       1          1        1         1

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 50  0 |  b = Iris-versicolor
  0  0 50 |  c = Iris-virginica
```

Confusion matrix in above result all 50 instances of all classes was correctly classified .The summary results shows that number of correctly classified instances is 150 and number of incorrectly classified instances is zero .Hence high accuracy is verified.

4. REP TREE

REPTree [4]is a fast decision tree learner. It uses information gain as splitting criteria to build a decision tree. REPTee can be pruned as well. Only numerical values can be sorted. Missing values are dealt with by splitting the corresponding instances into pieces.

Learning algorithm[5]: iterative procedure is as follows

1. Label 'D' data points and form clusters.
2. Find best split using say gini measure.
3. Let 'X' be the attribute with greatest gini gain.
4. Let 'Q' be corresponding best split set .
5. Partition 'D' into D1 and D2 based on split attribute 'X' and label node as 'T'.
6. Create children nodes T1 and T2 for T.
7. Build the tree(T1,D1) and tree(T2,D2).

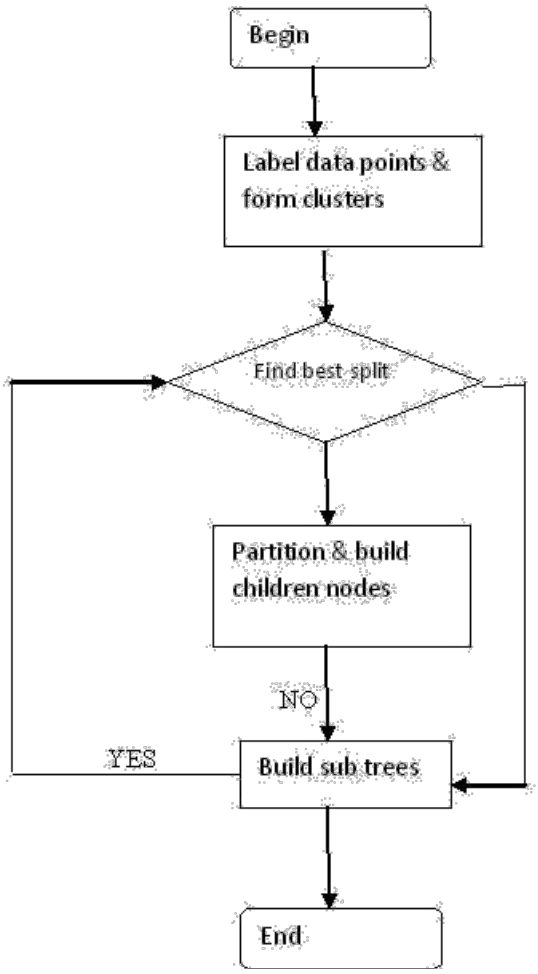

Figure3:Flow chart for learning algorithm

International Journal of Computer Network and Security(IJCNS)
Vol5. No.1 – Jan-March 2013 Pp. 12-15
©gopalax Journals, Singapore
available at : www.ijcns.com

Advantages of REPTree[6]:

- It manages both contiguous and categorical values.
- Easier to understand complex relationship between variables.
- Minimizes the effect of incorrect or missing values in final representation of tree.

Disadvantages of REPTree:

- Tree is unstable even for small changes in input data.
- Large tree models are difficult to analyze.

Experimental results for IRIS dataset in WEKA:

The training dataset was run using REPTree classifier to obtain following result:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         141               94      %
Incorrectly Classified Instances         9                6      %
Kappa statistic                          0.91
Mean absolute error                      0.0563
Root mean squared error                  0.1936
Relative absolute error                 12.6749 %
Root relative squared error             41.0599 %
Coverage of cases (0.95 level)          96.6667 %
Mean rel. region size (0.95 level)      41.3333 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               1         0         1          1         1          1        1         1         Iris-setosa
               0.92      0.05      0.902      0.92      0.911      0.866    0.948     0.886     Iris-versicolor
               0.9       0.04      0.918      0.9       0.909      0.864    0.948     0.871     Iris-virginica
Weighted Avg.  0.94      0.03      0.94       0.94      0.94       0.91     0.965     0.919

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 46  4 |  b = Iris-versicolor
  0  5 45 |  c = Iris-virginica
```

Confusion matrix shows that for class ‘a’ all 50 instances were correctly classified. For class ’b’ 46 out of 50 instances were correctly classified. For class ‘c’ 45 out of 50 instances were correctly classified. The summary shows that 141 instances out of 150 were classified. Hence showing lesser accuracy than random forest.

Comparison of random forest and REPtree[7]:

Table . 1 Comparison of classifiers

Random forest	REPTree
1. High accuracy of classification	1. Lesser accuracy of classification.
2.works well for large dataset	2. works well for small dataset.
3.can not handle complex relationships among variables.	3.can handle complex relationship among variables
4.computation of prototypes which gives relationship between variables and classification	4. difficult to construct prototypes.

5. CONCLUSION

Classification is a field of pattern recognition which creates models to group similar data. Classification trees help us to group expected data and hence draw conclusions from results. Random forest is one such classifier which classifies data labels based on number of votes and data is sampled independently. RepTree fast regression tree learner which sorts only numerical values.we have presented learning algorithm and the flowchart for above classifiers. The pros and cons of above classifiers is explicated. Experiments were conducted using the open source tool WEKA 3.7 and results are examined for IRIS dataset and conclusions are drawn. Using the experimental results we verified higher accuracy of random forest over REPTree.

REFERENCES

1. The Elements of Statistical Learning: Data Mining, Inference, and Prediction
   By Trevor Hastie, Robert J. Tibshirani, J. Jerome H. Friedman
2. http://www.stat.berkeley.edu/~breiman/RandomForests/
3. http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf
4. http://www.cs.cornell.edu/people/dobra/papers/secret-extended.pdf
5. http://www.knime.org/files/nodedetails/weka_classifiers_trees_REPTree.html
6. Design and Analysis of Randomized Algorithms: Introduction to Design Paradigms
   By J. Hromkovic
7. http://www.cs.manchester.ac.uk/resources/library/3rd-year-projects/2011/Akhil.Sharma.pdf