**COMPARATIVE EVALUATION OF MACHINE LEARNING ALGORITHMS**

**CS6000 RESEARCH METHODOLOGIES**

**Gayathiri Shriram (ms5754)**

**CSU East Bay**

**ABSTRACT**

This research paper represents the comparative evaluation of machine learning algorithms. I have chosen the decision tree induction from supervised learning category and K-means clustering from unsupervised learning category for this research paper. The paper gives an overview of machine learning, explains each of the learning techniques in detail along with its algorithms. Also, an application implementing both the algorithms is compared and evaluated. The algorithms are implemented using WEKA software and results of the two machine learning algorithms are compared and evaluated. Research is conducted on a weather forecasting application to warn about natural disasters mainly focused on earthquakes. The forecasting is done with both the learning techniques and ideas are presented based on the findings.

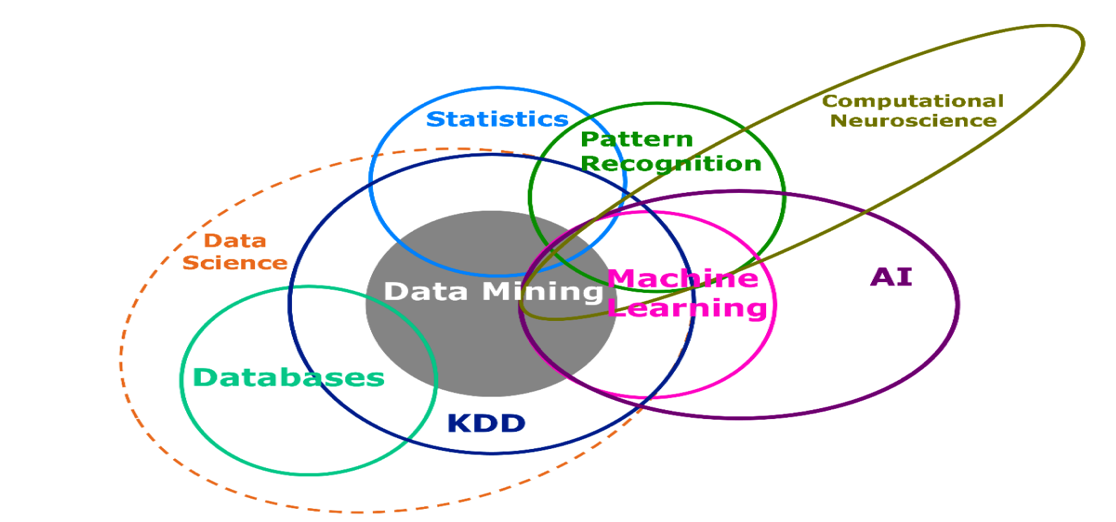
**INTRODUCTION**

Machine learning is the science of getting computers to act without being explicitly programmed.

It is a branch of artificial intelligence that is concerned with building systems that require minimal human intervention in order to learn from data and make accurate predictions. It is the subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. It explores construction and study algorithms that can learn and make prediction on the data.

Machine learning is so pervasive today that people use it many times a day without knowing they do as it is closely related to many fields. Also it is an assumption among the researches that machine learning is the best way to make progress towards AI.

The breakthroughs in psychology, mathematics and biophysics laid the foundations for machine simulation. Machine learning today is a mathematically rigorous discipline that encompasses sophisticated modeling, optimization, and learning research; it has concrete applications in medicine, software, robotics, and traditional business problems. Particularly in the business problem domain, there is significant overlap among the fields of data science, data mining, and machine learning. The below figure depicts the multidisciplinary nature of machine learning areas.

Figure 1

Machine learning is used in Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications.

There are many fields and application-using machine learning below are few widely publicized application using machine learning.

* Self driving Google car
* Amazon and Netflix are machine-learning applications of everyday life.
* Twitter – knowing what customers are saying about twitter – machine learning combined with linguistic rule.
* Fraud detection.

How is machine learning used today? Many of our day-to-day activities are powered by machine learning. Below listed are few such examples

* Fraud detection.
* Web search results.
* Real-time ads on web pages and mobile devices.
* Text-based sentiment analysis.
* Credit scoring and next best offers.
* Prediction of equipment failures.
* New pricing models.
* Network intrusion detection.
* Pattern and image recognition.
* Email spam filtering.

**WHY AND WHEN TO SELECT MACHINE LEARNING**

Selecting machine learning for an application is tricky and there are three conditions to be considered before finalizing on machine learning algorithms to be used for an application simulation

**Representation:**

A Classifier must be represented in some formal language that the computer can understand. The classifier set should in the hypothesis space for the leaner to learn.

**Evaluation:**

An Evaluation function is required to distinguish between good and a bad classifier.

**Optimization:**

The method to search among the classifiers should be of highest scoring one. The choice of optimization technique is key to the efficiency of the learner.

**MACHINE LEARNING METHODS**

Two of the most widely used machine learning methods are supervised and unsupervised learning. Most of machine learning is supervised learning. Semi supervised and reinforcement learning is two other technologies that are used sometimes.

**Supervised Learning:**

Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known. The functions are inferred from labeled training data. Training data consists of training examples. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances.

In order to use and solve a problem using supervised learning, the below steps are followed.

* Determine the type of training examples.
* Gather a training set.
* Determine the input feature representation of the learned function.
* Determine the learning algorithm to use.
* Complete the design.
* Evaluate the accuracy of the learned function.

The basic learning process of supervised machine learning is depicted in the below diagram.

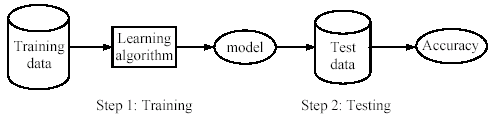


Figure 2

**Unsupervised Learning:**

Unsupervised data is used against the data that has no historical tables. This learning technique seems much harder the goal is to have the computer learn how to do something that we don't tell it how to do. Since the examples given to the learner are unlabeled there is no error or reward signal to evaluate the solution. Unsupervised learning is important since it is likely to be much more common in the brain than supervised learning. For instance there are around 106 photoreceptors in each eye whose activities are constantly changing with the visual world and which provide all the information that is available to indicate what objects there are in the world, how they are presented, what the lighting conditions are, etc.

**WEKA**

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real world

data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset.

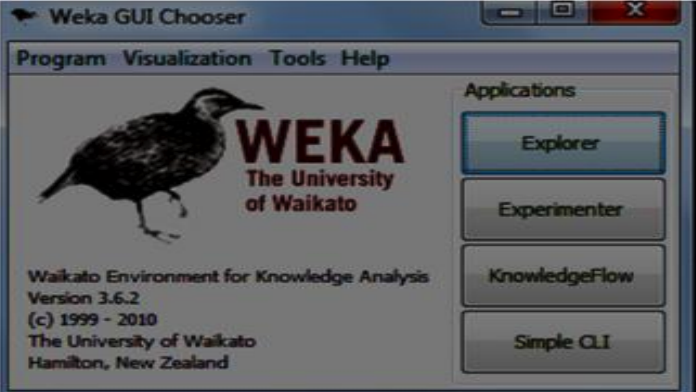
****

Figure 3

**LEARNING ALGORITHMS**

The learning algorithms compared in this paper are

1. Decision Tree Induction
2. K-means Clustering

**Decision Tree Induction**

Decision tree induction is a supervised learning algorithm. It is one of the most widely used techniques for classification. It uses decision tree as predictive model that maps the observation about an item to item’s target value.

A decision tree consists of nodes, edges and leaves.

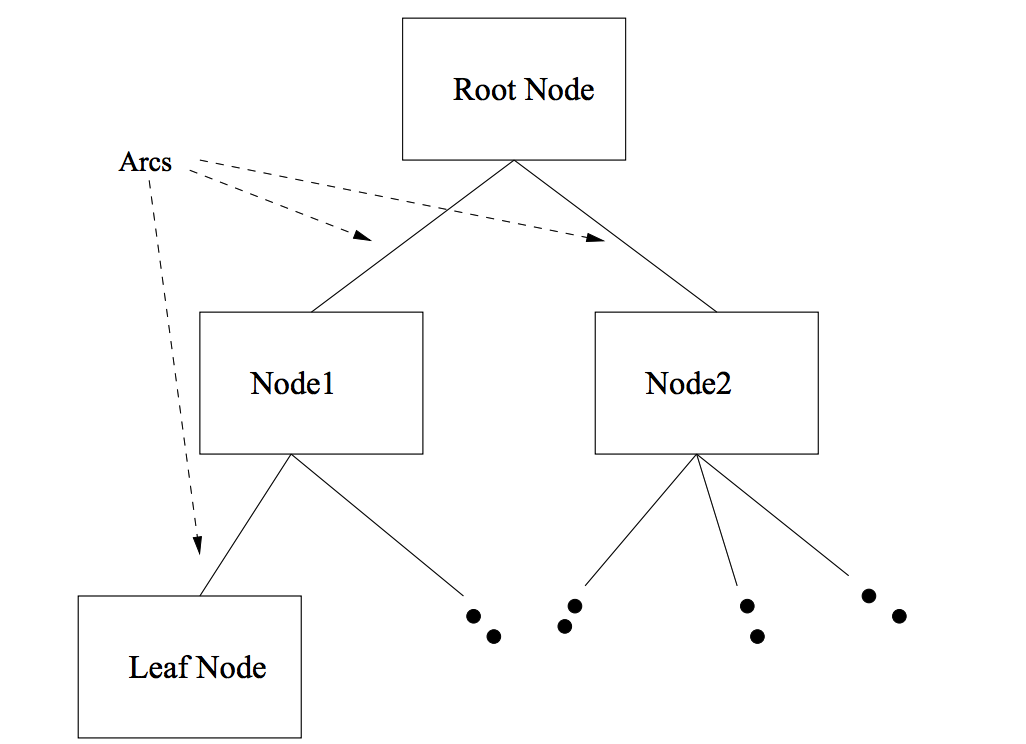
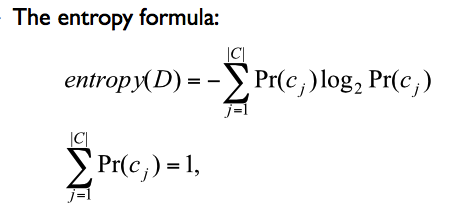


Figure 4

There are many popular decision tree algorithms and this paper deals with ID3 (**Iterative Dichotomiser 3)**  and its successor C4.5.

**The main ideas behind the ID3 algorithm are:**

* Each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. A leaf node corresponds to the expected value of the output attribute when the path from the root node to that leaf node describes the input attributes.
* In a “good” decision tree, each non-leaf node should correspond to the input attribute which is the most informative about the output attribute amongst all the input attributes not yet considered in the path from the root node to that node. This is because we would like to predict the output attribute using the smallest possible number of questions on average.
* Entropy is used to determine how informative a particular input attribute is about the output attribute for a subset of the training data. Below entropy formula is used for the calculation.



**C4.5 Algorithm**

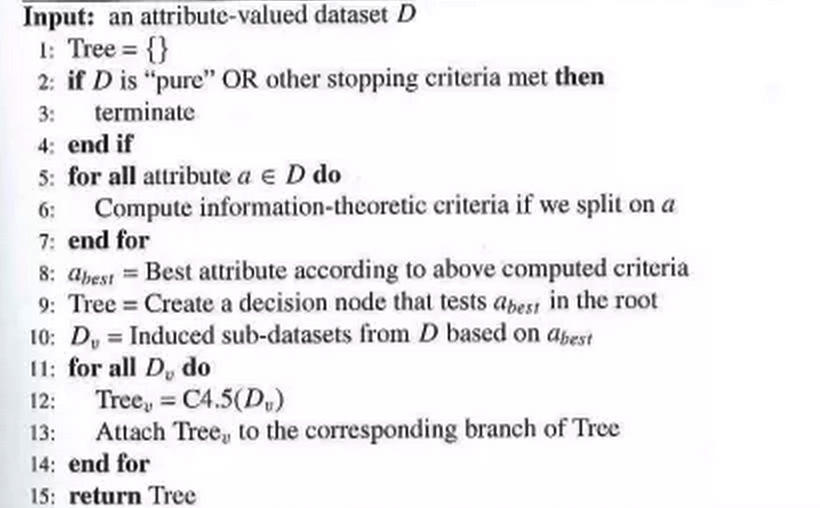
C4.5 decision tree learning algorithm is built the same way as ID3 using the concept of entropy.

The algorithm has few base cases

* All the samples in the list belong to the same class.
  + When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
* None of the features provide any information gain.
  + In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
* Instance of previously unseen class encountered.
  + Again, C4.5 creates a decision node higher up the tree using the expected value

**Pseudo Code:**

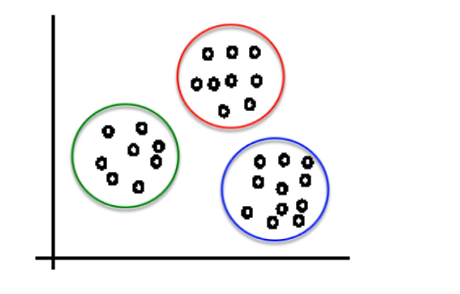
1. Check if algorithm satisfies termination criteria
2. Computer information-theoretic criteria for all attributes
3. Choose best attribute according to the information-theoretic criteria
4. Create a decision node based on the best attribute in step 3
5. Induce (i.e. split) the dataset based on newly created decision node in step 4
6. For all sub-dataset in step 5, call C4.5 algorithm to get a sub-tree (recursive call)
7. Attach the tree obtained in step 6 to the decision node in step 4
8. Return tree



**K-Means Clustering:**

Clustering is a technique for finding similarity groups in data, called clusters. It is often called unsupervised learning**.**

Sample clustering

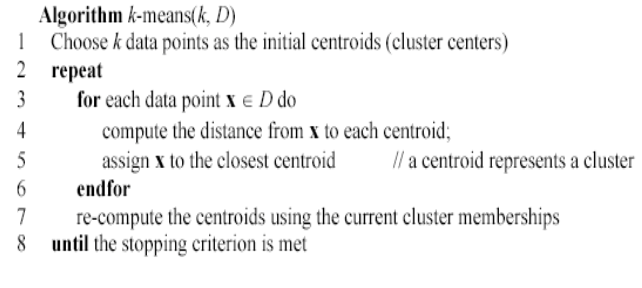
Figure 5

**K-means Algorithms:**

Given k, the k-means algorithm works as follows:

1. Randomly choose k data points (seeds) to be the initial centroids, cluster centers
2. Assign each data point to the closest centroid
3. Re-compute the centroids using the current cluster memberships.
4. If a convergence criterion is not met, go to (II).

**Pseudo Code:**

****

**COMPARATIVE EVALUATION OF ALGORITHMS**

As part of the research, outcomes of the decision tree and clustering algorithms are compared and evaluated for a **weather forecasting application to warn about natural disasters like earthquakes.**

The two algorithms are implemented using **WEKA (**Waikato Environment for Knowledge Analysis

) popular machine learning software written in Java

**Decision Tree Implementation:**

Using WEKA , C4.5 algorithm for data set containing details about earthquake is implemented

**Data Set**

@relation weather.earthquake

@attribute richterscale {0, 2.9, 3.5, 4, 5, 6.9, 7, 8.5, 9}

@attribute description {small, moderate, major, great, super}

@attribute occurrence {daily, monthly, yearly, rarely}

@attribute movement {small, moderatesudden, strongsudden, severesudden, extreme}

@attribute safe {yes, no, precaution}

@data

0,small,daily,small,yes

2.9,small,yearly,small,precaution

3.5,moderate,yearly,moderatesudden,precaution

4,moderate,monthly,strongsudden,no

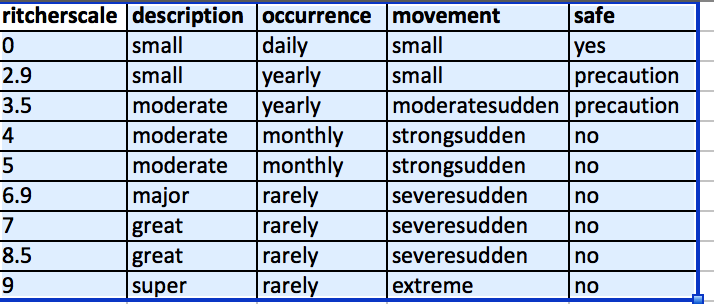
5,moderate,monthly,strongsudden,no

6.9,major,rarely,severesudden,no

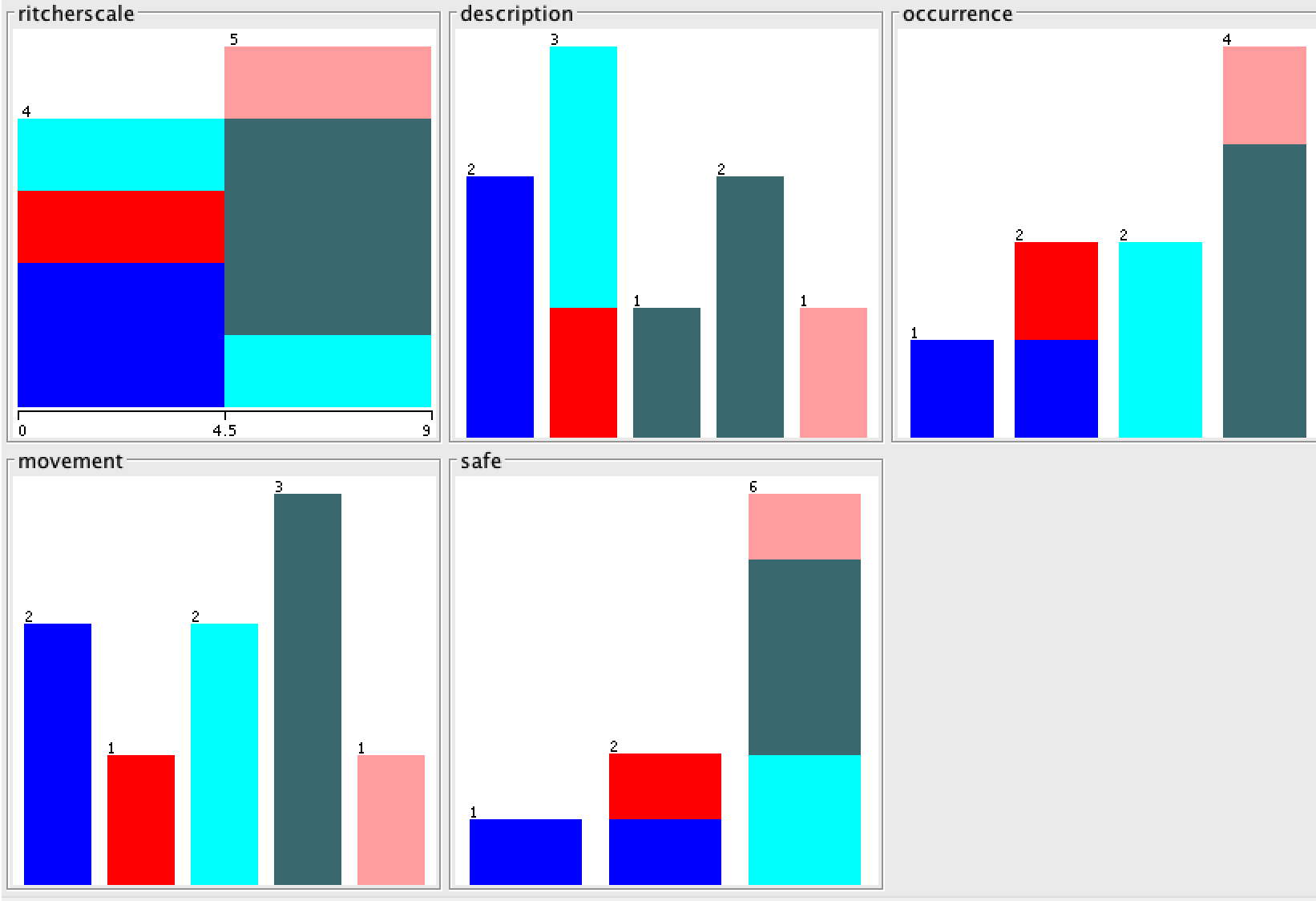
7,great,rarely,severesudden,no

8.5,great,rarely,severesudden,no

9,super,rarely,extreme,no

****

Visual Representation based on movement



**Implementation output of ID3(J48 in WEKA)**

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: weather.earthquake

Instances: 9

Attributes: 5

ritcherscale

description

occurrence

movement

safe

Test mode:evaluate on training data

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

J48 pruned tree

------------------

ritcherscale <= 3.5: precaution (3.0/1.0)

ritcherscale > 3.5: no (6.0)

Number of Leaves : 2

Size of the tree : 3

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances 8 88.8889 %

Incorrectly Classified Instances 1 11.1111 %

Kappa statistic 0.7692

Mean absolute error 0.0988

Root mean squared error 0.2222

Relative absolute error 27.5862 %

Root relative squared error 54.1828 %

Total Number of Instances 9

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0 0 0 0 0.875 yes

1 0.143 0.667 1 0.8 0.929 precaution

1 0 1 1 1 1 no

Weighted Avg. 0.889 0.032 0.815 0.889 0.844 0.97

=== Confusion Matrix ===

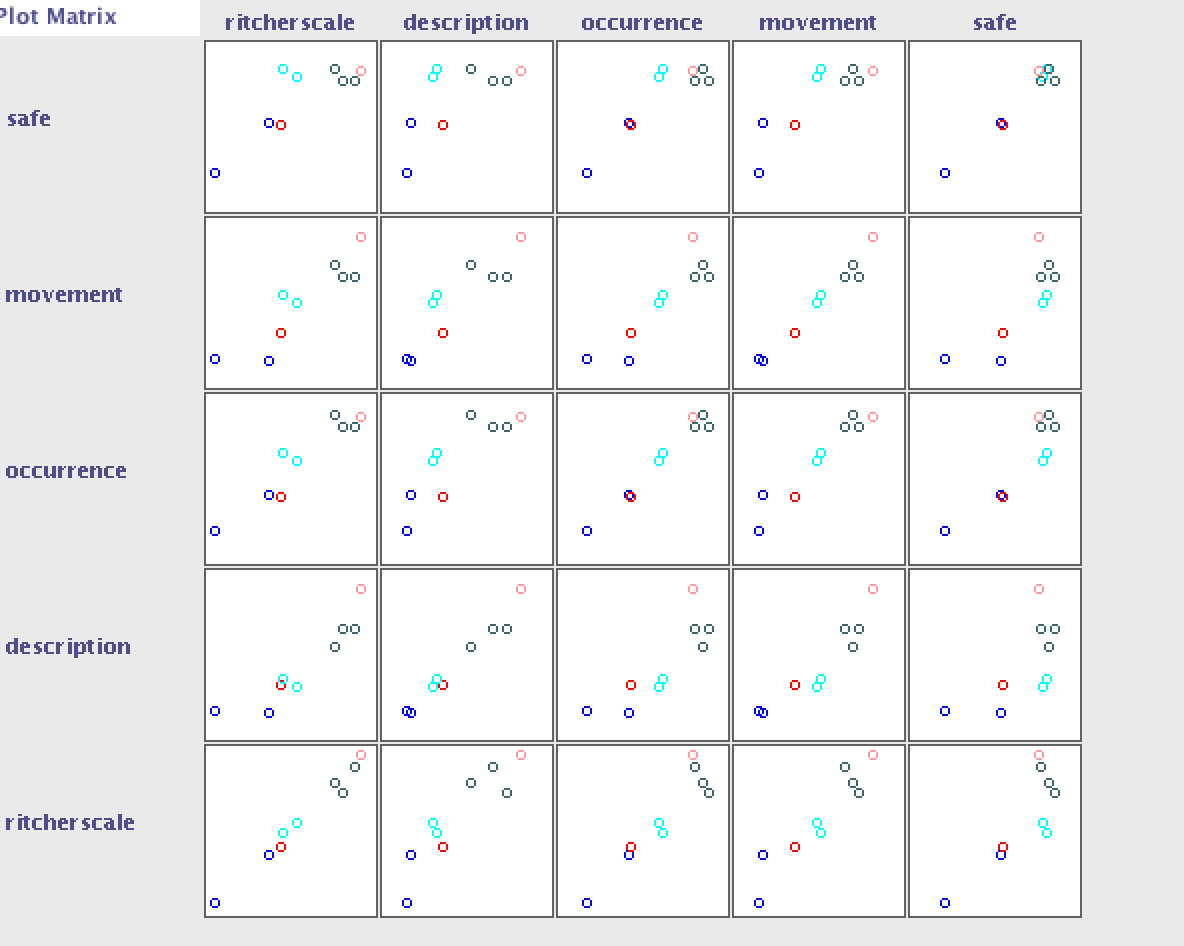
a b c <-- classified as

0 1 0 | a = yes

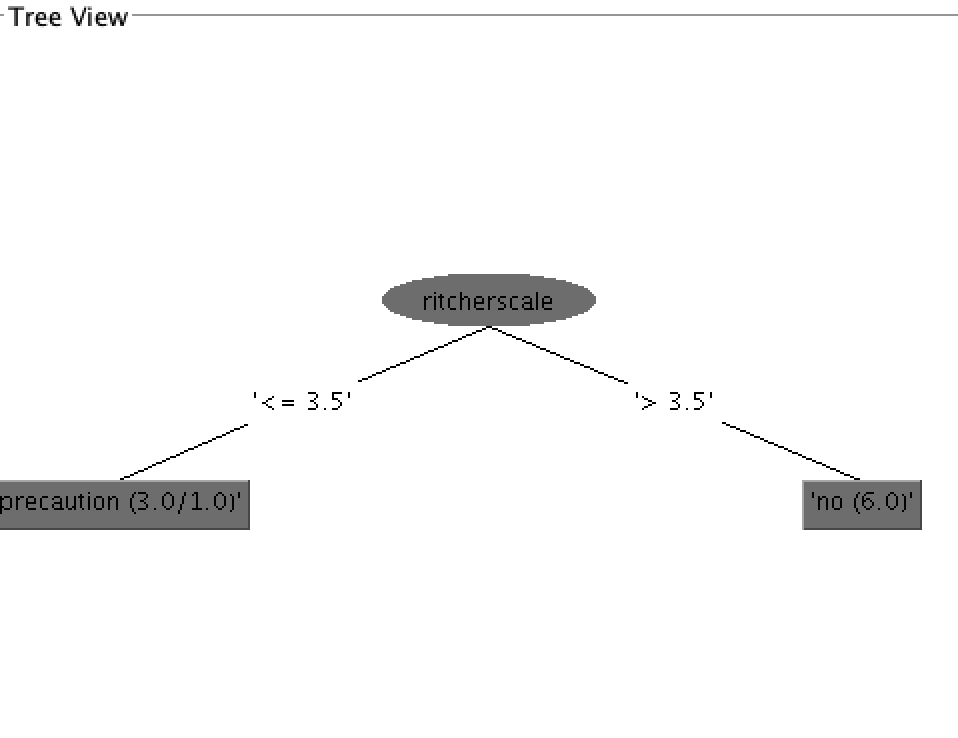
0 2 0 | b = precaution

0 0 6 | c = no

**Plot Matrix**

****

**Tree Visualization**



**K-means Clustering Implementation**

Using WEKA , simple K-means clustering algorithm for data set containing details about earthquake is implemented.

**Implementation Output of K-means Clustering**

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: weather.earthquake

Instances: 9

Attributes: 5

ritcherscale

description

occurrence

movement

safe

Test mode:evaluate on training data

=== Model and evaluation on training set ===

kMeans

======

Number of iterations: 2

Within cluster sum of squared errors: 14.395052910052911

Missing values globally replaced with mean/mode

Cluster centroids:

Cluster#

Attribute Full Data 0 1

(9) (2) (7)

========================================================

ritcherscale 5.2 1.45 6.2714

description moderate small moderate

occurrence rarely daily rarely

movement severesudden small severesudden

safe no yes no

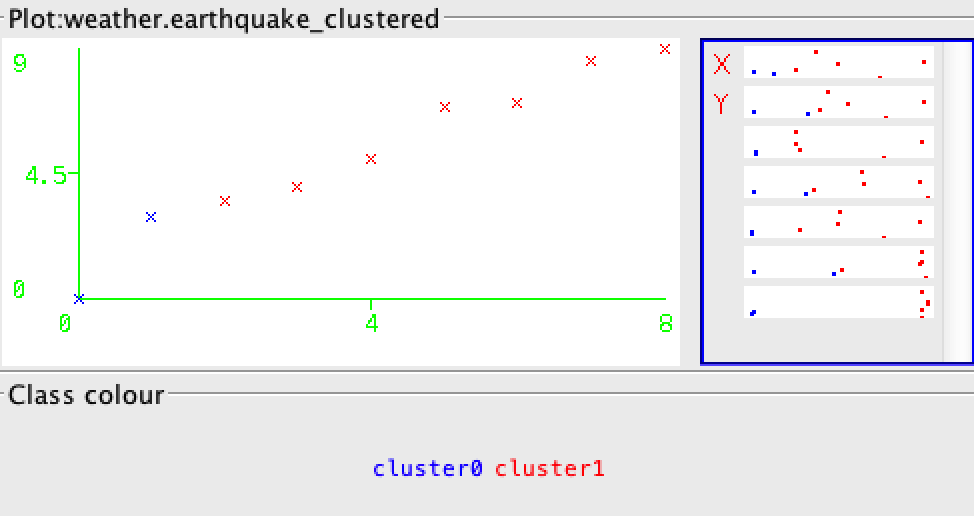
Time taken to build model (full training data) : 0 seconds

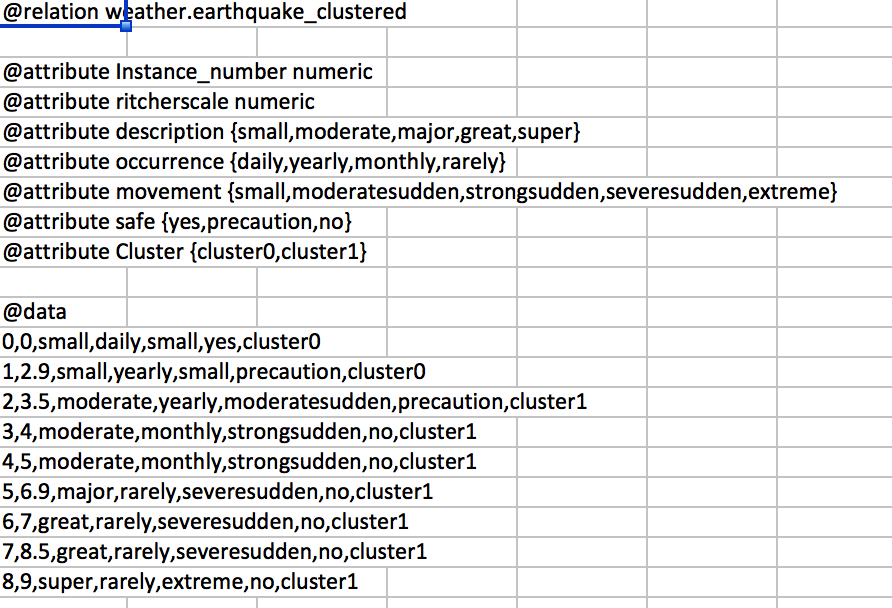
=== Model and evaluation on training set ===

Clustered Instances

0 2 ( 22%)

1 7 ( 78%)

****



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster0** | | | | |
| **ritcherscale** | **description** | **occurrence** | **movement** | **safe** |
| 0 | small | daily | small | yes |
| 2.9 | small | yearly | small | precaution |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster1** | | | | |
| **ritcherscale** | **description** | **occurrence** | **movement** | **safe** |
| 3.5 | moderate | yearly | moderatesudden | precaution |
| 4 | moderate | monthly | strongsudden | no |
| 5 | moderate | monthly | strongsudden | no |
| 6.9 | major | rarely | severesudden | no |
| 7 | great | rarely | severesudden | no |
| 8.5 | great | rarely | severesudden | no |
| 9 | super | rarely | extreme | no |

**Comparative Analysis**

Percentage analysis of decision tree learning algorithms

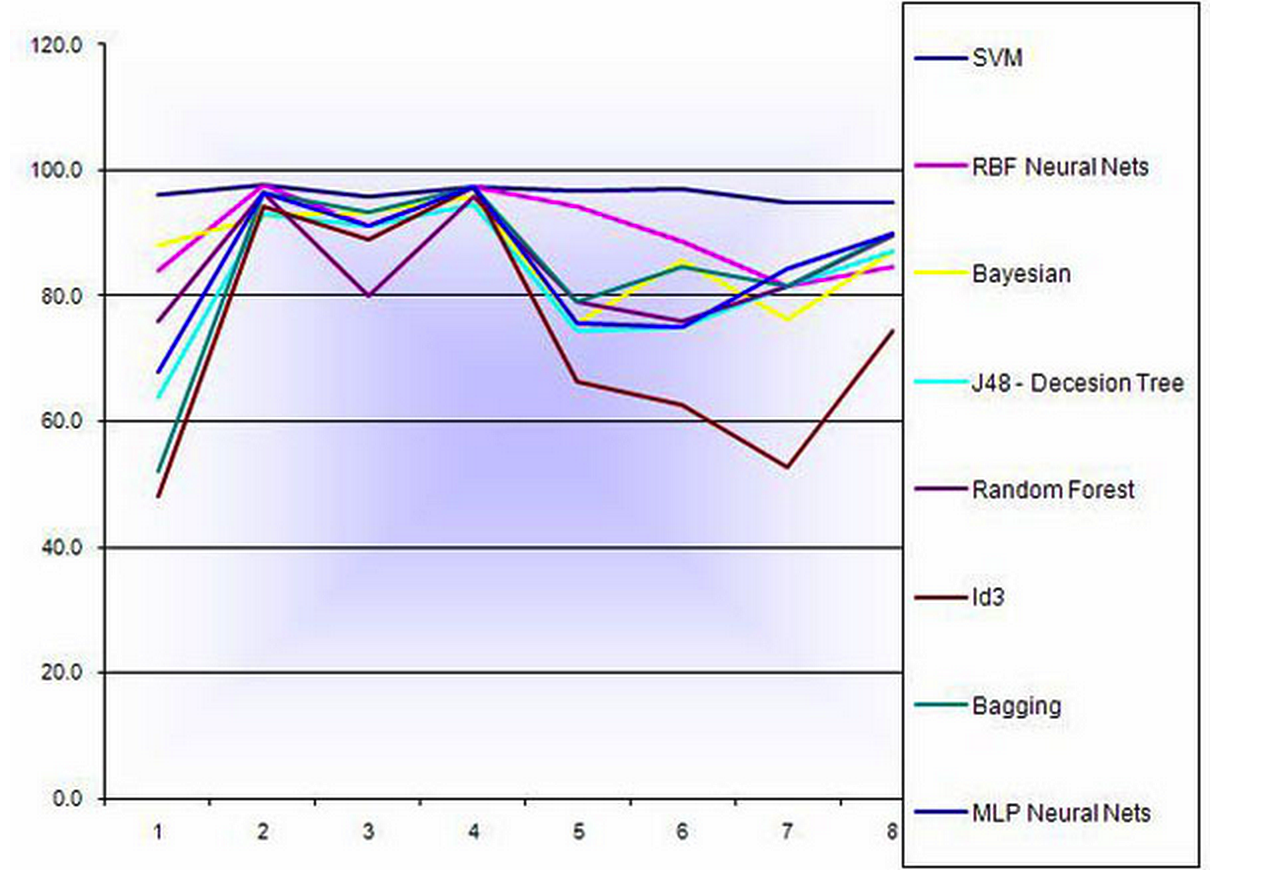
****

Figure 6

Percentage analysis of clustering algorithms

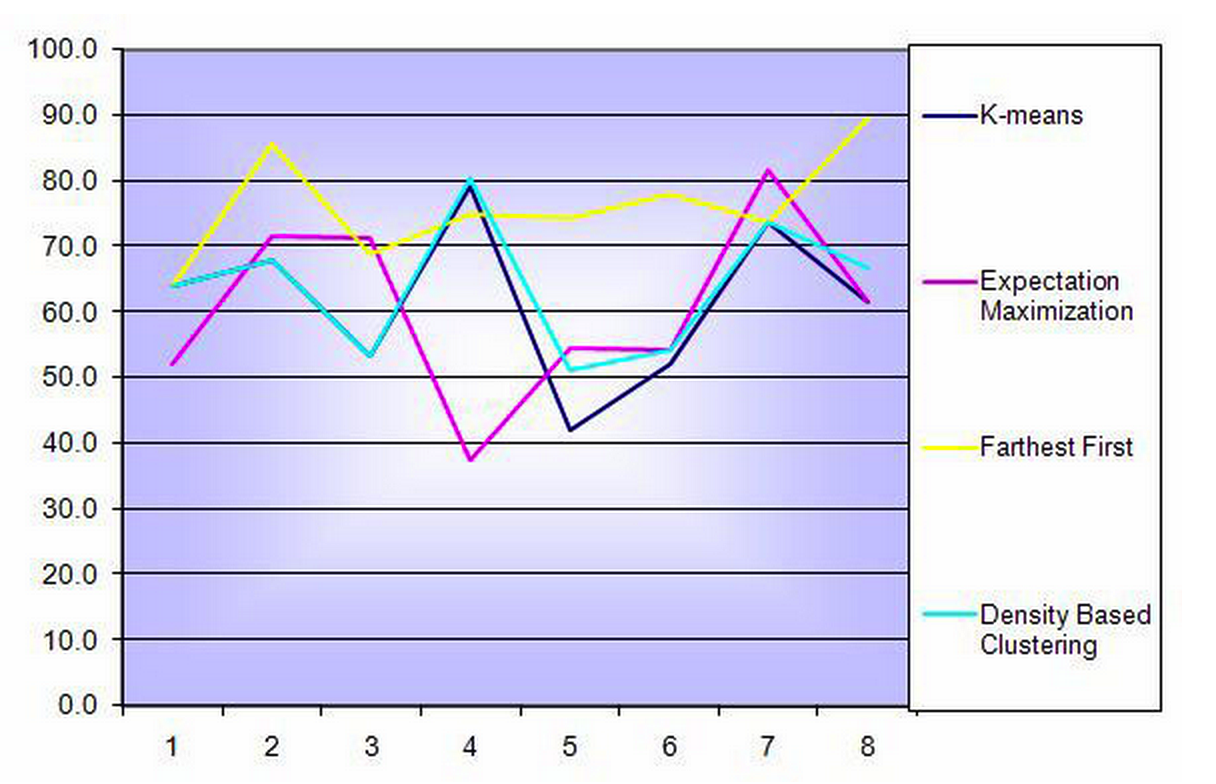


Figure 7

**Results of Decision Tree learnt from the research**

* Decision tree are simple and easy to understand and interpret as we can see from the tree graph we retrieved from WEKA
* Requires little data preparation
* Handle both numerical and categorical data
* Possible to validate a model using statistical tests
* Performs well with small dataset.
* In addition to the advantages it has some disadvantage as well
* Decision tree cannot be optimum. Also they are based on each node hence cannot be relied upon to provide global optimum solution
* Some concepts like XOR, parity are not well explained in decision trees.
* Time consumption is more as training a model and using it to make decisions for large data set.

**Results of K-means Clustering learnt from the research**

* Easy to implement and it is faster
* We get to set the number of clusters we need which makes the decision easier
* The clusters are very clear and provide a much clearer picture compared to decision tree.
* K-means provide tighter clusters when global data is considered.
* Time consumption is very less as no training is required and data set are fed directly into the algorithm.

**CONCLUSION**

This paper has provided a comparison of decision tree and clustering algorithms used in forecasting natural disasters. At this point in research clustering seems to be more appropriate on global front of data. Both the clustering as well decision tree has disadvantages. In decision tree the analysis goes node by node hence when huge data volume is considered decision might not provide appropriate solution. On the other hand clustering algorithms does not handle ambiguous data well. However it is early to conclude that this is the optimum algorithm to use. Further research in this area by improvising the data set to include more details we can come up best optimum analysis to be used.

**REFERENCE**

* Figure 1 – Web Data Mining by Bing Liu
* Figure 2 – Web Data Mining by Bing Liu
* Figure 3 – Web Data Mining by Bing Liu
* Figure 4 – Web Data Mining by Bing Liu
* Figure 5 – http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2386055/
* Figure 6 – http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2386055/
* Web Data Mining –Bing Liu
* http://en.wikipedia.org/wiki/Supervised\_learning
* http://en.wikipedia.org/wiki/Decision\_tree\_learning
* <http://gisgeography.com/100-earth-remote-sensing-applications-uses/>
* <http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm>
* <http://www.cs.waikato.ac.nz/ml/weka/>
* <https://weka.wikispaces.com/Frequently+Asked+Questions>
* <http://en.wikipedia.org/wiki/Weka_%28machine_learning%29>
* <http://www.sms-tsunami-warning.com/pages/richter-scale#.VWy2KVxViko>
* <http://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html>