New approach for Dynamic File Classification Solution Using Decision Tree Algorithm

Francis Bambo [#1], Deepak Pathak[*2], Gagan Sharma[#3]

Department of Computer Science and Engineering, Sri Satya Sai College of Engineering, Bhopal

[1]francisbambo1@gmail.com

[2] deep_325@yahoo.com

[3] gagansharma.cs@gmail.com

Abstract - Static file classification approach was been used for several year to carry out the process of the classification of the file in many industrial and business world, but due to the huge and increasing in the file classification operations, it became un sufficient to provide a good quality and became very slow to execute a very huge number of files in the a short time.

Dynamic file classification came to exist to overcome this problem, as a solution to this problem many dynamic software ware release. DFCS research is one of the best solutions to automate the files in the system to be processed further on dynamically.

In this paper we have use DFCS as the best solution to overcome static file classification problem in a very wide range of use. DFCS applies Decision Tree Algorithm researching, IF THEN rule based as well as many techniques we have used to improve our classification performance.

The result of our solution was amazing, and very good.

Keywords —DFCS, Decision tree, file classification

I. INTRODUCTION

A. FILE CLASSIFICATION:

File Classification can be defined as the process of putting together a similar files from a different folders of the system into a group of similar folders, based on their features and attributes. Using supersized learning approach of the machine learning algorithms to learn and predict the targeted values from the dataset, based on given attributes.

The model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret. [1]File classification is the process of putting files into categories based on their features and properties. Files can be classified either manually or automatically. Files can be classified either manually or automatically.

Manual classification of large numbers of Files has the following problems associated with it:

It is power consuming, needs a lot of classification specialist and It is time consuming.

Classification quality reduces by the use of the traditional (statically) manual approaches. [1] (Hastie TJ).

B. *Decision Tree Algorithm*

As it is stated from the title of the algorithm, the algorithm takes the assigned dataset, and generates a tree of multi-nodes for its final prediction decision, using learning approach from a given attributes of the dataset in two paths of execution.[3] (Patel N).

The dataset is assigned to the first node of the tree (root node) and the algorithm initiate a model to learn from the attributes given to the existed dataset. The model uses the first execution path resulted data to formulate its prediction values decision.

This kind of algorithms can solve both statistical (Regression – classification) problems. [10] (Loh W)

How it is works? And what principals this algorithm uses? The execution processes? Creation of the algorithm tree? Part of the algorithm tree? ... etc., we will explain and cover all these detail in coming unites of our research.

II. PROBLEM FORMULATION

A. **Problem Formulation: The purpose of this Thesis is to build a module and design a software application system to deal with a balk huge files identifying, classifying and predicting its attributes correctly, removing duplication copies of the file from the system predicted list, to demonstrate high performance and the good accuracy of the best Dynamic file classification approach.**

B. **Approach Used: In our paper, we proposed a new file classification approach to solve the problem of manual file classification. We designed a most reliable and user-friendly file classification system**

based on Decision Tree, IF THEN, Language Integrated Query (LINQ), in a supervised learning mining technique, which organized the files by different classification types. DFCS contains two levels:

1- **One level classification: classifying the files by only one type at time.**
2- **Two level classification: files are classified by many types at a time.**

We used MD5 encoding algorithms to find out the duplicated files in the system Thus, the MD5 helps in reducing the time of comparing as well as the space of the system.

III. Literature Review

3.1 Overview of Decision Tree Algorithms:

[10] (Loh W) as it is stated from the title of the algorithm, the algorithm takes the assigned dataset, and generates a tree of multi-nodes for its final prediction decision, using learning approach from a given attributes of the dataset in two paths of execution.

First execution path (learning):

The dataset is assigned to the first node of the tree (root node) and the algorithm initiate a model to learn from the attributes given to the existed dataset.

Second execution path (Predicting): the model uses the first execution path resulted data to formulate its prediction values decision.

This kind of approach can solve both statistical (Regression – classification) problems.

How to use the algorithm:

We should assign the dataset with its selected attributes to our first node in our algorithm tree to start its first execution path as we have explained it above to give us a good result of predicted values. Then second execution path start from where the first path stops to continue comparing these attributes with our objective variables.

This execution repeat itself for the sub-nodes in the tree till final decision achieved.

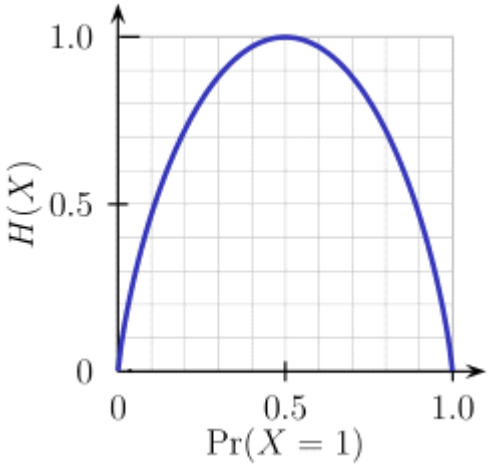
3.2 Attribute Selection Measures:

Decision trees uses a lot of measures to decide which attributes to be considered as appropriate attributes to be placed in the root node or which of these attributes to be distributed to the rest of the nodes in the tree as sub-nodes.

Here are some attribute measurements being used by DT algorithm:

3.3. Entropy Measurement:

Entropy is a testing rate of unordered or irregular information being executed in the moment. The greater the entropy implies the difficulty of reaching the end of this random information. The example for this is a deicing a coin.



This graphical representation of Entropy algorithm shows that, when the curve starts from the point zero and going in increase up till certain point then it start decrease down to a certain point greater than zero started point

Zero point implies the leaf node in id3 algorithm and greater than zero implies need for more division of the node

Mathematically:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

(S) represents the Current state.

(P) represents Probability of an event (i) of state S or Percentage of class (i) in a node of state S.

Mathematically as:

$$E(T,X)=\sum_{c\in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

where T= Current state and X = Selected attribute

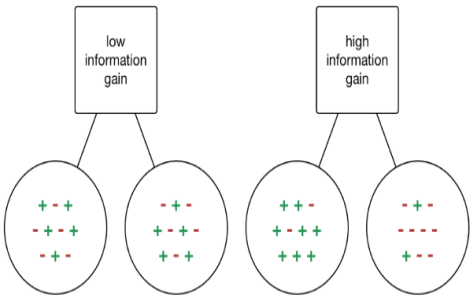
[21][22] (Sagal M)

These criterions will compute values for every attribute, the attribute with a greatest value is selected to be the root.

3.4 Information Gain Attribute measurement:

This algorithm is used to carry a test on the training samples to prove the efficiency of the attributes selected, based on the objective classification values,

The objectives of this algorithm is to provide and create a (DT) tree with a greatest (IG) Information Gain and less (E) Entropy. It has overcame the problems that ID3 suffered from.



As shown by the above diagram, this algorithm is intended to minimize the entropy measure values by computing and comparing the difference between entropy values the whole dataset (before) and (after) segregation and partition, according to the given attribute values base.

Final result after the algorithm calculations and execution paths, the small partitions of the datasets with different values are most likely accepted by the algorithm than other partition.

Mathematically:

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{K} Entropy(j, after)$$

(before) = dataset before it been partitioned.

K = subsets generated after partition took place.

(j) = subset j after the partition of the dataset.

3.5 Gini Index (success p or failure q) Measurement:

Can be defined as cost function used to evaluate segregation in the dataset. Gini index is computed by subtracting the sum of the squared probabilities of each class from one.

Larger partitions and easy to execute whereas information gain prefer smaller partitions with distinct values

If the value is more than similarity expected to be greater and powerful.

Mathematically:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

(P^2+Q^2)

IV. Result and Decisions

4.1 Technologies Used:

- **LINQ** technique is used to go through directory and folders instead of traditional iteration which also helps in reducing the time of moving through the folders and files.
- **Thread mechanism** is used to make the system work in asynchronous manner (Tasks are working together) which helps in speeding our project.
- **Background worker** is used to make some tasks work in the system background which means, there is no need to wait for one task to complete in order to achieve another task.
- **Md5 algorithms** is used to find out the finger print of the files which are duplicated.

4.2 Screens:

Figure 4.2.1 Login form

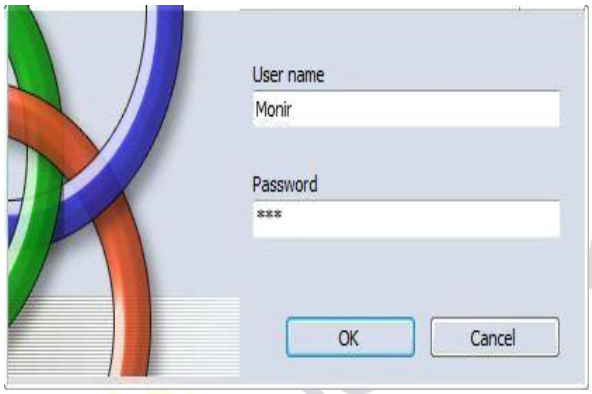


Figure 4.2.2 Main form

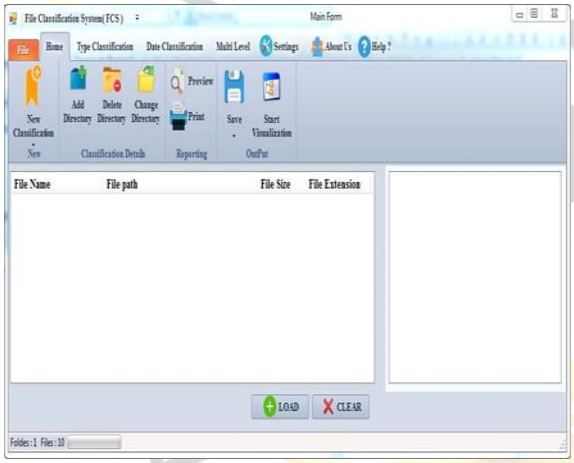


Figure 4.2.3 General Classification

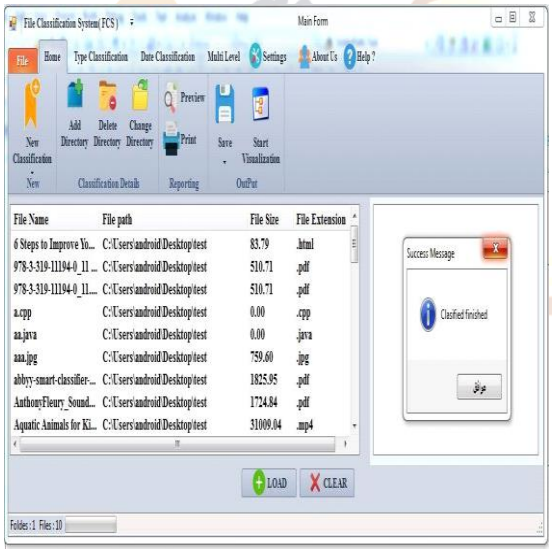


Figure 4.2.4 General classification out put

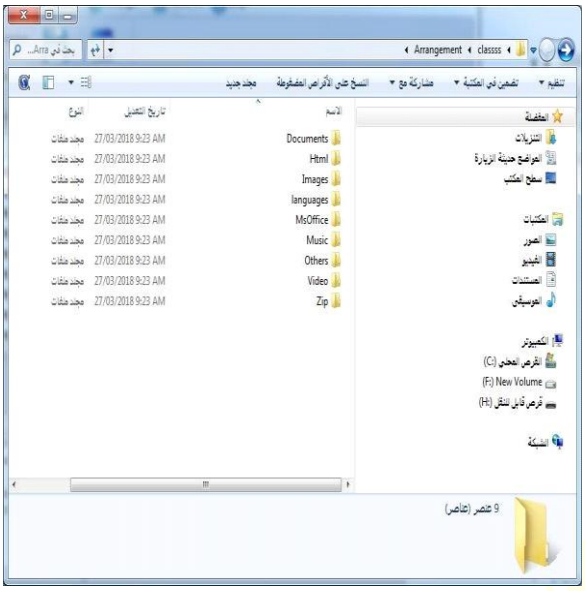
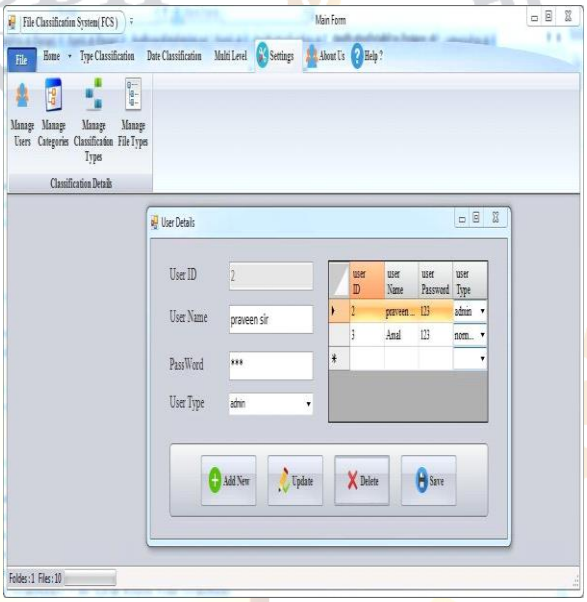


Figure 4.2.5 Manage users:



Reference

[1]. Hastie TJ, Tibshirani RJ, Friedman JH. The Elements of Statistical Learning: Data Mining Inference and Prediction. Second Edition.

[2]. Fallon B, Ma J, Allan K, Pillhofer M, Opportunities for prevention and intervention with young children

[3]. Patel N, Upadhyay S. Study of various decision tree pruning methods with their empirical comparison in WEKA.

[4]. Berry MJA, Linoff G. Mastering Data Mining: The Art and Science of Customer Relationship Management.

[5]. Hastie T, Tibshirani R, Friedman J.The Elements of Statistical Learning.

[6]. Zibran MF. Department of Computer Science. Diagnostic and Statistical

Manual of Mental Disorders – Fourth Edition.

[7]. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees.

[8]. Quinlan RJ. C4.5: Programs for Machine Learning.

[9]. Kass GV. Anexploratory technique for investigating large quantities of categorical data.

[10]. Loh W, Shih Y. Split selection methods for classification trees.

[11]. Bhukya DP, Ramachandram S. Decision tree induction-an approach for data classification using AVL–Tree.

[12]. Lin N, Noe D, He X. Tree-based methods and their applications: Handbook of Engineering Statistics.

[13]. SAS Institute Inc. SAS Enterprise Miner12.1 Reference Help, Second Edition.

[14]. IBM Corporation. IBM SPSS Modeler 17 Modeling Nodes.

[15]. Is See5/C5.0 Better Than C4.5? Australia: Rulequest Research

[16]. IBM Corporation. IBM SPSS Statistics 23 Command Syntax Reference.

[17]. Batterham PJ, Christensen H, Mackinnon AJ. Modifiable risk factors predicting major depressive disorder at four-year follow-up: a decision tree approach.