International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

Elimination of redundant Files using Feature Selection Algorithm

Ch Sundeep[1], B Vamshi [2], P Sampath[3], V Ethirajulu , B. Tech, M. Tech[4]

[1]Student, Dept. of Computer Science Engineering SRM University, Tamilnadu, India
[2]Student, Dept. of Computer Science Engineering SRM University, Tamilnadu, India
[3]Student, Dept. of Computer Science Engineering SRM University, Tamilnadu, India
[4]Assistant Professor, Dept. of Computer Science Engineering SRM University, Tamilnadu, India

---

Abstract – *Demonstrating late advances in the machine learning systems to best in class discrete decision models, we build up an way to deal with oversee reason the captivating and complex fundamental activity system of a manager (DM), which is portrayed by the DM's needs and attitudinal character, near to the properties investment, to give a couple of delineations. This work presents a different method with regards to any learning of proximity relations.*

Key Words: **Data Mining**

1. INTRODUCTION

The process of data mining is to assess data from several instances and summarize it into relevant information in order for us to understand it completely. The techniques of data mining are the result of incessant research methods. This idea started when workable data was actually stored on computers, continued with significant advancements in accessing data, and more recently, generating technologies which permit users to steer through their data in actual time scenarios. It consists of five major parts:

- ETL (Extract, transform, and load) onto the database system with your data.
- Store and handle the data in a multidimensional environment.
- Provide complete access to data to authenticated persons.
- Examine the data by any software application.
- Express the data in a widely used form that is easily understandable.

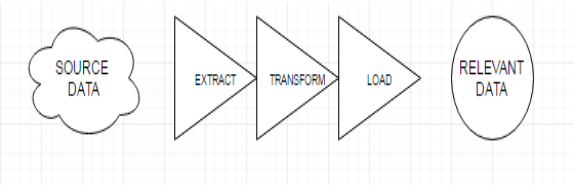

Fig – 1: ETL Diagram

2. METHODOLOGY

In this work, we propose a simple yet robust mechanism to eliminate redundant files which are uploaded to the Database every day. We use Feature Selection algorithm to sort the files based on whether or not they are redundant. The algorithm also reduces the Dimensionality of Data to a manageable level. The most important step here is the Deduplication process which does the actual work of eliminating the redundant files.

The process involves several steps such as: Creation of a File, uploading the File, Deduplication, Storing the File, and Removing the File as shown in Fig – 2.

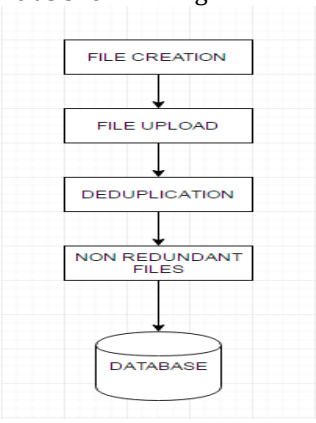

Fig – 2: Block Diagram

2.1 FILE CREATION

The File containing the Data must be uploaded onto the Database. The Uploaded File can be of any type such as .DOC, .TXT and .PDF. The Data to be uploaded should have a maximum size of 50 megabytes.

2.2 FILE UPLOAD

Once the File is created, it is uploaded onto the system and checked whether it is redundant or not. If a File with the same Data already exists, then the system should show a message about its existence and must not accept the new file.

3. DEDUPLICATION

The Deduplication step is the most important task to determine whether a given File is redundant or not. This step takes the Uploaded File as input and checks whether it is redundant or not. It uses Feature selection Algorithm to do just that.

Feature selections are also known as variable selection or attribute selection. It is the choice of attributes in the data that are really important to the predictive modeling complication you need to work on.

Feature selection is not to be confused with dimensionality reduction. Usually both methods help decrease the attribute count in a dataset. Reduction of dimensionality mostly develops new samples of attributes, where as feature selection keeps and removes the attributes generated by not altering them at all.

1: Input: candidate feature x^j, j = 1; 2; . . . ; J, the label $l(x_i)$, i = 1; 2; . . .,N

2: Initialize the selected feature subset F_0 = [], and the weight subset a_0 = [], and the distribution $w^1{}_{i,k} = 2 / (n^2 - n)$

3: Compute the pair wise dissimilarities $d(x^j{}_i, x^j{}_k)$ for all candidate features

4: t = 0

5: while the stop condition is not satisfied, where the stopping condition is t is equals to T_x or $Err(F_t, a_t) > Err(F_{t-1}, a_{t-1})$

do

6: for j , all feature indices without being selected **do**

7: Build the binary classifier for each feature. $Hj^t = sign(v(x_j) - d(x^j{}_i, x^j{}_k)$

8: Evaluate the error rate of binary classifier corresponding each feature $error(j) = \sum_I \sum_{k, k != i} w^t{}_{I,k}[(I - l(x_{ij})^j = (x^j{}_k)] - a_j . h^j{}_t)^2$

9: Find the best binary classifier $J^* = \arg\min_j error(j)$; and obtain the weight a^* , and threshold $v(x^{j*})$

10: end for

11: Update the selected feature subset and weight subset $F^t = [F_{t-1}, x^{j*}]$, $a_t = (a_{t-1}, a^{j*})$

12: Compute the error of the strong classifier $err(F_t; a_t)$.

13: Update the distribution $W_{i,k}{}^{t+1} = W_{i,k}{}^{t+1} / \sum_I \sum_{k, k != I} W_{i,k}{}^{t+1}$

14: t = t + 1:

15: end while

16: Output: F_t and a_t:

The three types of feature selection algorithms are as follows:

- filter method
- wrapper method
- embedded method

3.1 FILTER METHOD

The Filter method maintains a ranking based system to allocate for the scoring of every feature. Its features are established correctly by the resultant score which might be accepted to be kept in the generated dataset. The methods are mostly a result of a single variation to consider the feature separately while regarding the variable as a dependant one.

3.2 WRAPPER METHOD

Wrapper method considers the usage of a unique set of attributes for any given search problem, while several permutations of it are designed, and measured to many other such sequences. A predictive sequence is used to assess this unique arrangement of features and designate a score depending on the correctness of the developed model.

The searching may be sequential such as a BFS type, it may also be having a random probability distribution such as hill-climbing methodology, or it can involve heuristics, like the different types of passes to append and discard features.

3.3 EMBEDDED METHOD

Embedded method contains features which usually help build and contribute to the correctness of the model while it is being developed by the system. The most perceived and common type of embedded method is a regularization method.
Regularization method is also often called as a penalization method that maintains additional set of constraints for making the most of a predictive algorithm (usually a regression algorithm) that shifts the entire developed model towards a conundrum of lower complexity (lesser number of coefficients).

4.0 NON REDUNDANT FILES AND DATABASE STORAGE

After the deduplication process, the non redundant files are separated from the redundant ones. They are then stored onto the Database which can be viewed later once the user logs in to the system. When the File is uploaded onto the system, it does not mean it gets stored in it.

5.0 CONCLUSION

In this paper, the Feature selection algorithm is used to eliminate the redundant Files from getting stored in the Database. A feature subset can generally give rise to an higher classification ability. We also know that the use of Feature Selection algorithm helps in the reduction of the complex issue of data dimensionality by a considerable amount which discards the unneeded variables and attributes from the process easily enough for a better and accurate assessment of the uploaded Data.

REFERENCES

[1] Rie Kubota Ando, Tong Zhang tzhang, “A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data

[2] Taiping Zhang, Pengfai ren, Yao Ge, Yuan tang, C,L, Philip Chen, “ Learning proximity relations for feature selection”