Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Issuing of Pollution Under Control Certificate using ID3 algorithm

Rupali Bhardwaj
CSE , Bahra university, India.

Sonia Vatta
CSE , Bahra university, India.

Abstract— Learning that is based on induction is the inductive learning. Decision tree algorithms are very famous in inductive learning. These kinds of algorithm use inductive methods for the appropriate classification of the objects with the given attributes. These algorithms are very beneficial in the classification of the objects and are mainly used in expert systems. In this paper the ID3 decision tree learning algorithm is used to find out whether there are any changes in the present decision rules for issuing of PUCC (Pollution under Control Certificate) when some new attributes are added. Here the three studies are done regarding the issuing of PUCC and in each study a new attribute is added to the dataset to get the decision rules as resultant. The algorithm is implemented in the java language.

Keywords— inductive,classification,ID3,decision learning, algorithm

I. INTRODUCTION

ID3, Iterative Dichotomiser 3 is a decision tree learning algorithm which is used for the classification of the objects with the iterative inductive approach. In this algorithm the top to down approach is used. The top node is called as the root node and others are the leaf nodes. So it's a traversing from root node to leaf nodes. Each node requires some test on the attributes which decide the level of the leaf nodes. These decision trees are mostly used for the decision making purpose [8]. Decision tree learning is a procedure for calculating the target value having discrete function. The function that has been learned is symbolized by a decision tree. For the inductive inference the decision tree learning is one of the most commonly and broadly used methods which are practical in nature [1][3].

The decision tree learning algorithms are mainly used because of the three reasons :

1. Decision tree is a good infer from the particular cases that are unobserved instance.
2. The calculations in these methods are efficient and are proportional to the instances that are observed.
3. At the final, the decision tree which is produced is easily understood by the human. [2][3]

II. CONCEPTS OF BASIC ID3

A. ID3 Basics

ID3 is simple decision tree learning algorithm which uses the greedy top to down search to build the tree which will decide the decision rules. For this there is a requirement for some mathematical concepts. The two concepts which are basically involved in ID3 are ENTROPY and INFORMATION GAIN.

B. Entropy

Entropy is used to calculate the randomness of each attribute. The formula for entropy is

ENTROPY(S)=-P(positive)log2P(positive)- P(negative)log2(-negative)

P(positive) are the positive examples in S
P(negative) are the negative examples in S
S is the sample training set having positive and negative values.
Entropy can also be said as the measure of the impurities in the training set that is used[3].
After calculating the entropy the decision tree which is constructed is very large.

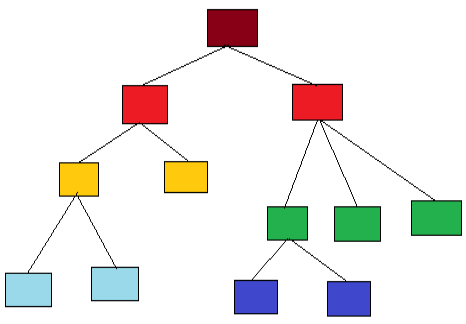

Figure 1: Decision tree after entropy is calculated.

So to reduce the depth of the decision tree we use the next concept that is Information Gain.

© 2013, IJARCSSE All Rights Reserved

C. *Information Gain*

In ID3 the splitting criteria of the node uses the Information Gain. The attribute having the highest gain will be considered as root node. The formula for IG is

GAIN(S,A)=Entropy(S)-Sum for V from 1 to n of (|Sv|/|S|)*Entropy(Sv)

As IG reduces the depth of the tree the following will be the resultant tree

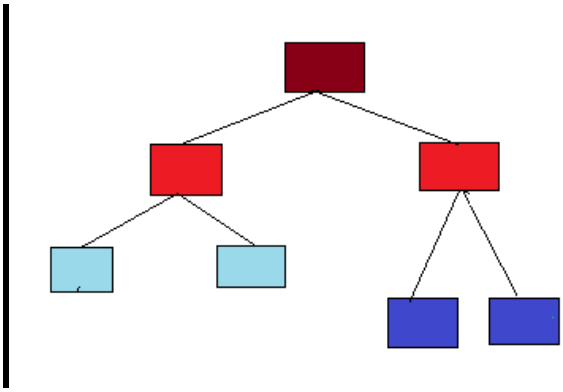

Figure 2: Decision tree with reduced depth after IG is calculated.

The procedure will be recursively used again and again until the required classification is not gained.

D. *Rules for Classifying.*

If the entropy of the attribute is 0, it is a homogeneous node and there is no need to classify further.

If the entropy of the attribute is 1, it is a heterogeneous node and there is a need to classify further.

III. LITERATURE REVIEW

The first work on ID3 was done by J.R Quinlan in 1986.He synthesizes decision trees that have been used in variety of systems and one such system he described as ID3[4]. Data mining techniques basically use the ID3 algorithm as it's the basic algorithm of classification. Some reusable components were identified from ID3 algorithm by milija suknovic (et.al). They combined the reusable components so as to allow the replication of original algorithm. They also found out that their modification will create a new induction algorithm[14]. An interesting implementation or the use of ID3 was done in the field of networks security. Victor H.Goreia (et.al) use ID3 to web attack detection. It outputs the decision rules which are easy to classify and grasp the root of an attack[9]. Another researcher was Sonika Tiwari who used the improved ID3 for detecting Network Anomalies with horizontal portioning based decision tree[10]. Data mining techniques basically use the ID3 algorithm as it's the basic algorithm of classification. In the medical field ID3 were mainly used for the data mining.Ruijuan Hu used the ID3 algorithm for retrieving the data for the breast cancer which is carried out for the primarily predicting the relationship between the recurrence and other attributes of breast cancer[7]. Prediction of common diseases in mobile phones and television is also done with the use of ID3 by L.Sathish Kumar and A.Padmapriya. They propose method phone which helps the people to know about the diseases to avoid the death rate and diseases affected people count[13]. Mary Slocum implemented the ID3 for data mining in the medical research. She modifies the large amount of data and transforms the data into information which can be used to make a decision[11].

The implementation and evaluation of ID3 was done by some authors on different examples, like Anand Bahety implemented the algorithm on the "Play Tennis" database. He classified whether the weather is suitable for playing tennis or not?[8]. T.Y HSU 662096093 did the survey on the simple "loan application" dataset. He classified whether a person will get a house loan or not?. Kumar Ashok (et.al) did the same on the "census 2011 of India" dataset. He classified it by using ID3 algorithm. Another research implemented the same to find out whether a person is sunburnt or not? [12]

In this paper an example on ID3 is implemented to know the rules to issue the PUCC to the vehicles. There are three datasets on which the ID3 algorithm is applied to get the decision rules. Query arises is whether the vehicle will get the PUCC?

IV. OBJECTIVE OF THE RESEARCH WORK

The main objectives of the research work here are

- First, to construct the decision tree until the appropriate classification is reached.
- Another is to generate the decision rules for the problem.

V. RESEARCH METHODOLOGY

The methodology used to reach to the conclusion of the research work is as follows.

The language in which the algorithm is implemented is the java language. The dataset used contains all the information about the problem (the attributes, their values etc). The dataset is further used for the construction of the decision tree. The three datasets are created for the three studies. Each dataset has one added attribute. Once the dataset is created then the entropy is calculated and after that the information gain is calculated for each dataset separately.

© 2013, IJARCSSE All Rights Reserved

The final step is the construction of the decision trees and the decision rules. After that the comparing of all the three studies is done to find out the change in the present scenario.

V. RESULTS AND DISCUSSION

The present scenario for the issuing of PUCC is the only manual check up to measure the pollution with the physical machine, and the database has only the values for the pollutants that are emitted by the vehicle. The issuing of PUCC only depends upon the manual check up till today.

Present Rule is:

IF manual check up =approved
THEN
Issue= Yes
ELSE
Issue=No

So approach of this paper is to change this rule by adding some more attributes which decide the issuing of PUCC. So these studies will decide whether there are attributes that can also be taken with the manual check up to decide whether to issue PUCC to the vehicle or not.

After pruning of many attributes related to the problem the final attributes for the datasets are

1. Fuel –having sub attributes
   A) Petrol
   B) Diesel
   C) CNG
2. Category – having sub attributes
   A) Two wheeler
   B) Three wheeler
   C) Four wheeler
3. Kilometers-the minimum distance travelled by any vehicle is kept 5000 km,so here two sub attributes are made which depict whether this kms are travelled or not.so the sub attributes are
   A) Covered
   B) Not covered
4. Service- the service of the vehicle would be done before three months from getting PUCC,so the sub attributes are
   A) Yes
   B) No
5. Year – according to the government any vehicle whose manufacturing year completes the 15 years will not be valid for the use of transport. So this can also be an important attribute to decide the issuance of the PUCC. It is having the sub attributes as
   A) Valid
   B) Non valid
6. Manual check up-having sub attributes
   A) Approved
   B) Not approved
7. Issue-this is the target attribute having the sub attributes as
   A) Yes
   B) No

Now the three studies are discussed and their Decision trees and decision rules are generated below.

STUDY 1

Table I

Name	Fuel	Category	Kilometers	MC	Issue
Riva	Petrol	Two	Not covered	Pass	Yes
Shobhna	Petrol	Two	Covered	Pass	Yes
Puneet	Petrol	Four	Covered	Fail	No
Rahul	Petrol	Four	Covered	Pass	Yes
Nishi	Diesel	Four	Covered	Fail	No
Surya	Diesel	Three	Covered	Fail	No
Rehan	Diesel	Three	Not covered	Pass	Yes
Sameer	CNG	Three	Not covered	Pass	Yes
Rashmi	Petrol	Four	covered	Fail	No

Here the kilometers is the attribute which is added to the present scenario. The decision tree and the decision rules for this study generated are

© 2013, IJARCSSE All Rights Reserved

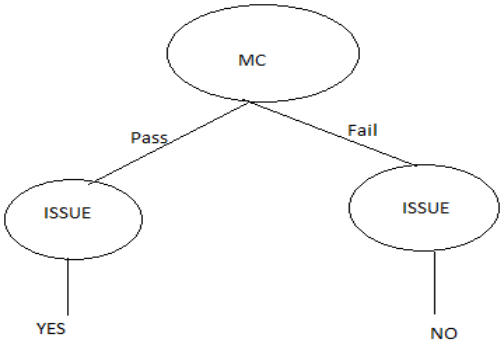

Figure 3: Decision tree for study 1

The rules generated after adding the kilometer attribute are

```
If (MC==”Pass”){
        Issue=”Yes”;
} elseif(MC==”Fail”) {
        Issue=”No”;
}
```

STUDY 2

Table II

Name	Fuel	Category	Kilometers	Service	MC	Issue
Riva	Petrol	Two	Not covered	No	Pass	Yes
Shobhna	Petrol	Two	Covered	Yes	Pass	Yes
Puneet	Petrol	Four	Covered	No	Fail	No
Rahul	Petrol	Four	Covered	Yes	Pass	Yes
Nishi	Diesel	Four	Covered	No	Fail	No
Surya	Diesel	Three	Covered	No	Fail	No
Rehan	Diesel	Three	Not covered	No	Pass	Yes
Sameer	CNG	Three	Not covered	No	Pass	Yes
Rashmi	Petrol	Four	covered	Yes	Fail	No

Here the service is the attribute which is added to the present scenario. The decision tree and the decision rules for this study generated are

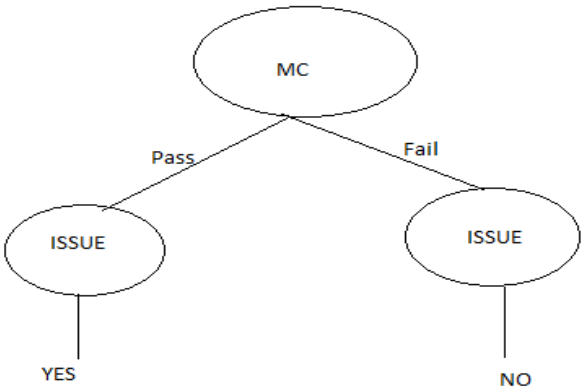

Figure 4: Decision tree for study 2

The rules generated after adding the kilometer attribute are

```
If (MC==”Pass”){
        Issue=”Yes”;
} elseif(MC==”Fail”) {
        Issue=”No”;
}
.
```

© 2013, IJARCSSE All Rights Reserved

STUDY 3

Table III

Name	Fuel	Category	Kilometers	Service	Year	MC	Issue
Riva	Petrol	Two	Not covered	No	Valid	Pass	Yes
Shobhna	Petrol	Two	Covered	Yes	Unvalid	Pass	No
Puneet	Petrol	Four	Covered	No	Valid	Fail	No
Rahul	Petrol	Four	Covered	Yes	Valid	Pass	Yes
Nishi	Diesel	Four	Covered	No	Unvalid	Fail	No
Surya	Diesel	Three	Covered	No	Valid	Fail	No
Rehan	Diesel	Three	Not covered	No	Valid	Pass	Yes
Sameer	CNG	Three	Not covered	No	Valid	Pass	Yes
Rashmi	Petrol	Four	covered	Yes	Unvalid	Fail	No

Here the year is the attribute which is added to the present scenario. The decision tree and the decision rules for this study generated are

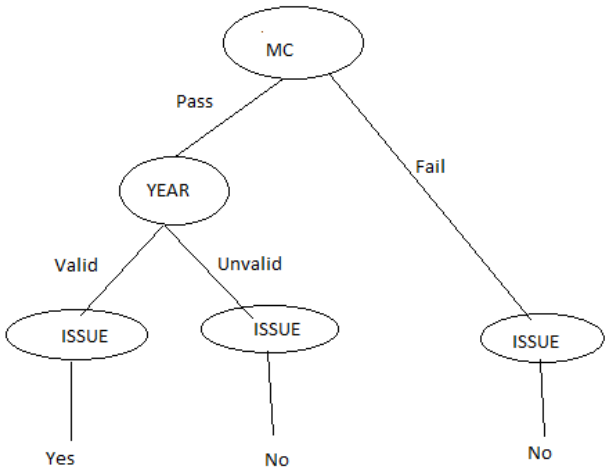

Figure 5: Decision tree for study 3

The rules generated after adding the kilometer attribute are

```
If (MC==”Pass”){
        If(Year==”Valid”){
                Issue=”Yes”;
} elseif(year==”Unvalid”){
                 Issue=”n”
}
 elseif(MC==”Fail”) {
                Issue=”No”;
}
```

VI. **Conclusion**

The studies and their implementation conducted here conclude that the decision tree learning algorithm ID3 works well on any classification problems having dataset with the discrete values [8]. Related to the research work it concludes that year is the attribute which is also taken with the manual check up for the issuing of PUCC to the vehicles.

So the new decision rules are

```
If (MC==”Pass”){
        If(Year==”Valid”){
                Issue=”Yes”;
} elseif(year==”Unvalid”){
                 Issue=”n”
}
 elseif(MC==”Fail”) {
                Issue=”No”;
}
```

© 2013, IJARCSSE All Rights Reserved

So now whenever a person will go forgetting the PUCC,his vehicle has to pass not only the manual check up but also has to pass the test of age of the vehicle. If his vehicle is not 15 years old and passes the manual check up then only he will get the PUCC.With this the pollution as well as traffic on the roads will be reduced to some extent.

VII. **Future Scope**

According to the research done till now ,this research work will be pursued at the higher level by knowing the hidden pollutants which are not measured yet in the present systems. Researchers can work on the sulphides and nitrides emitted by the vehicle to reduce the pollution by creating some new machine or software.

References

[1] Tom M. Mitchell, (1997), *Machine Learning*, Singapore, McGraw-Hill.
[2] Paul E. Utgoff and Carla E. Brodley, (1990), *'An Incremental Method for Finding Multivariate Splits for Decision Trees'*, Machine Learning:Proceedings of the Seventh International Conference, (pp.58),Palo Alto, CA: Morgan Kaufmann.
[3] Wei Peng, Juhua Chen and Haiping Zhou, of ID3,*' An Implementation Decision Tree Learning Algorithm'*,University of New South Wales, School of Computer Science & Engineering,Sydney, NSW 2032, Australia.
[4] Quinlan, J.R. 1986, *Induction of Decision trees*, Machine Learning.
[5] http://www.cs.cornell.edu/Courses/cs578/2003fa/missing_featsel_lecture.ppt
[6] http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm
[7] Ruijuan Hu,(2011),*' Medical Data Mining Based on Decision Tree Algorithm',* Computer and Information Science,Vol. 4, No. 5; September 2011, www.ccsenet.org/cis.
[8] Anand Bahety,*' Extension and Evaluation of ID3 – Decision Tree Algorithm'*.,University of Maryland, College Park.
[9] Vıctor.H.Garcıa, Raul Monroy and Maricela Quintana,' Web Attack Detection Using ID3'. http://homepage.cem.itesm.mx/raulm/pub/ifip06.pdf.
[10] Sonika Tiwari and Prof. Roopali Soni,*' Horizontal partitioning ID3 algorithm A new approach of detecting network anomalies using decision tree'*, International Journal of Engineering Research & Technology (IJERT)Vol. 1 Issue 7, September – 2012.
[11] Mary Slocum,',*Decision making using ID3*,RIVIER ACADEMIC JOURNAL, VOLUME 8, NUMBER 2, FALL 2012
[12] Kumar Ashok, Taneja H C, Chitkara Ashok K and Kumar Vikas,*' Classification of Census Using Information Theoretic Measure Based ID3 Algorithm'* . Int. Journal of Math. Analysis, Vol. 6, 2012, no. 51, 2511 – 2518.
[13] L. Sathish Kumar and A. Padmapriya,*' Prediction for Common Disease using ID3 Algorithm in Mobile Phone and Television'*. International Journal of Computer Applications (0975 – 8887) Volume 50 – No.4, July 2012
[14] M. Suknovic et al(2011).*' Reusable components in decision tree induction Algorithms'*.
[15] Herbert Schildt, *Java 2: The Complete Reference*,5th edition, McGraw-Hill/Osborne
[16] http://www.wikipedia.com/
[17] http://www.google.com/

© 2013, IJARCSSE All Rights Reserved