# Induction of Decision Trees

S.G.CLARO

09/30/2018

## 1 Relivance

This article seems to match with the specific research on decision trees. It's composed by TDIDT family (Top Down Induction of Decision Trees) and Induction Tasks. There's alson an exemple of algorithm ID3. It's OK.

## 2 Quality

### 2.1 Where ?

1986 Kluwer Academic Publishers, Boston. Seems legit.

### 2.2 Who ?

J.R. Quinlan. Wikipedia :

> "John Ross Quinlan is a computer science researcher in data mining and decision theory. He has contributed extensively to the development of decision tree algorithms, including inventing the canonical C4.5 and ID3 algorithms. He also contributed to early ILP literature with First Order Inductive Learner (FOIL). He is currently running the company RuleQuest Research which he founded in 1997."
> Famous scientist in decision theory with a lot of publications.

### 2.3 When ?

1987. Pretty old.

### 2.4 Cited ?

Cited : 19849 times. Trustworthy.

### 2.5 Relevant ?

This article is at the base of decision trees theory. It's totally relevant !

# 3 Article Review

## 3.1 Introduction

The author explain the importance of IA in 1950's which permit to machine learning to become a central research area. That is to say, "learning provides a potential methodology for building high- performance systems." (J.R. Quinlan, 1987).There are different subfields listed in learning research and the author explain the basics of "knowledge-based expert system". Then he'll focus on one of these system family : TDIDT family (Top Down Induction of Decision Trees) and its algorithm example.

## 3.2 Main question

J.R.Quinlan tries to explain the importance of machine learning in creating based-knowledge expert system. What are the mechanism behind ? What is the power and the secrets of one learning algorithm ?

## 3.3 Side questions/ Hypothesis

What is TDIDT family ? How do they work ? Limits ? Benefits ?

## 3.4 Plan approach

First Quinlan will outline the members and characteritics of TDIDT family, in inducing decision trees. Then he 'll develop the ID3 system, "that enable to cope with noisy and incomplete information"(Quinlan,1987). And improvement will be explained.

## 3.5 Methods

### 3.5.1 TDIT Family

- 3 dimensions in machine learning systems : learning strategies, representation of knowledge acquired, application domain.
- "they address all involve classification"(Quinlan,1987).
- Classification tasks : diagnosis of medical condition from symptoms, determining the game-theoric value of chess position, decide of weather.
- "The member of this family are sharply characterized by their representation of acquired knowledge as decision tress. [...] TDIT family are considerably less complex [...] it is still possible to generate knowledge in the form of decision trees that is capable of solving difficult problems of practical significiance."(Quinlan, 1987).

- Incremental methods needs an order, which the objects are presented. List of thoses different systems : MARVIN(Sammut,1985), Winston(1975), BACON(Langley, Bradshaw and Simon, 1983), INDUCE(Michalski, 1980). They develop decision trees for classification tasks.

- "These trees are constructed beginning with the root of the tree and proceeding down to its leaves."(Quinlan, 1987).

- Decision trees are expressed in terms of set of attributes or properties. They could come from a database of a patient's medical observations.

- Family tree of TDIDT system :

  - **CLS** : Hunt's Concept Learning System framework. Patriarch of the family."CLS construct a decision tree that attempts to minimize cost of classifying an object.

    > "This cost has component of two types : the measurement cost of determining the value of property A exhibited by the object, and the missclassification cost of deciding that the object belongs to the class J when its real class is K.CLS uses a lookahead strategy similar to minimax. At each stage, CLS explores the space of possible decision trees to a fixed depth, chooses an action to minimize cost in this limited space, then moves one level down in the tree."(Quinlan,1987).

  - **ID3** : Quinlan. Series of program developed from CLS. Chess challenged with pattern recognition. Form a decision tree iteratively from a window till all objects in the training set are all classified.

  - **ACLS** : (Paterson and Niblett, 1983) generalization of ID3. It permits properties that have unrestricted interger values. Used in image recognition (Shepherd,1983).

  - **ASSISTANT** : (Kononenko, Bratko and Roskar, 1984) ID3 as direct ancestor.Generalizes on the integer tributes of ACLS by permitting attributes with continuous values(real). It's not iterative but finer division than disjoint class. Used in medical domain

  - Others : ACLS commercials derivative. Westinghouse Electric's Water Reactor Division example. Incorporate user-friendly innovations and utilities.

### 3.5.2 Induction task

- Universe of objects that described in terms of a collection of attributes. Attributes = measure of important feature of an object, limited to taking set of discrete, mutally exclusive values.

- Example : Saturday Morning = Object and Classification task : Weather. So different attributes with values like outlook -¿ sunny, overcast or rain.

- Zeroth-order language : to characterize objects in the universe (state 0 of a Saturday Morning). How it seems to be usually.

- "Each object in the universe belongs to one of a set of mutually exclusive classes." (Quinlan, 1987).

- P and N : objects class. Two -class induction tasks : P for positive instances and N for negative isntances.

- Need of a training set of objects whose class is known. Induction task will develop a classification rule that can determine the class of any object from values of the attributes.

- Attribute inadequate : case when two attributes from different objects are the same. Not suited for induction task.

- Classification rule will be expressed as a decision tree. (see example table 1 training dataset and trees figure).

- Leaves = class names, other nodes = attrivute-based tests with branch for each possible case. Process until encounter leaf, then it gives class named.

- Essence of induction : move beyond the training set and predict other unseen objects. There are many correct decision trees.

- Case when we get two decisions trees, correct for training set, simpler is always the best. (other exemple figure).

- Methodology can be enhanced to deal with information noisy and/or incomplete.